

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau



INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

WO 98/12354

(51) International Patent Classification 6 : C12Q 1/68, C07H 21/04, G06F 17/00		A1	(11) International Publication Number: WO 98/12354
(21) International Application Number: PCT/US97/17002		(43) International Publication Date: 26 March 1998 (26.03.98)	
(22) International Filing Date: 18 September 1997 (18.09.97)		(74) Agents: SMITH, William, M. et al.; Townsend and Townsend and Crew LLP, 8th floor, Two Embarcadero Center, San Francisco, CA 94111 (US).	
(30) Priority Data: 60/025,740 19 September 1996 (19.09.96) US		(81) Designated States: AU, CA, JP, US, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).	
(60) Parent Application or Grant (63) Related by Continuation US 60/025,740 (CON) Filed on 19 September 1996 (19.09.96)		Published <i>With international search report.</i>	
(71) Applicant (for all designated States except US): AFFYMETRIX, INC. [US/US]; 3380 Central Expressway, Santa Clara, CA 95051 (US).			
(72) Inventors; and (75) Inventors/Applicants (for US only): LOCKHART, David, J. [US/US]; 610 Mountain View Avenue, Mountain View, CA 94041 (US). WONG, Gordon [US/US]; 239 Clark Road, Brookline, MA 02146 (US). SAFER, Penina [US/US]; 31 Kenwood Avenue, Newton, MA 02159 (US).			
(54) Title: IDENTIFICATION OF MOLECULAR SEQUENCE SIGNATURES AND METHODS INVOLVING THE SAME			
(57) Abstract <p>Novel means and methods for analyzing hybridization data derived from hybridization assays between a target nucleic acid and differently sequenced polynucleotide probes involve selecting probe sets that define reference sequences for sequence signatures and deriving useful data about the nature of the target nucleic acid molecule based on its hybridization to the probes. The methods are useful for determining whether the target contains a nucleic acid or polypeptide or sequence signature, whether the target encodes a member of a gene family, or whether the target is derived from one of any number of genes.</p>			

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece	ML	Mali	TR	Turkey
BG	Bulgaria	HU	Hungary	MN	Mongolia	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MR	Mauritania	UA	Ukraine
BR	Brazil	IL	Israel	MW	Malawi	UG	Uganda
BY	Belarus	IS	Iceland	MX	Mexico	US	United States of America
CA	Canada	IT	Italy	NE	Niger	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NL	Netherlands	VN	Viet Nam
CG	Congo	KE	Kenya	NO	Norway	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NZ	New Zealand	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	PL	Poland		
CM	Cameroon	KR	Republic of Korea	PT	Portugal		
CN	China	KZ	Kazakstan	RO	Romania		
CU	Cuba	LC	Saint Lucia	RU	Russian Federation		
CZ	Czech Republic	LI	Liechtenstein	SD	Sudan		
DE	Germany	LK	Sri Lanka	SE	Sweden		
DK	Denmark	LR	Liberia	SG	Singapore		
EE	Estonia						

5 **IDENTIFICATION OF MOLECULAR SEQUENCE SIGNATURES**
AND METHODS INVOLVING THE SAME

10 This application claims the benefit of the filing
date of United States provisional patent application
60/025,740, filed September 19, 1996, which is incorporated
herein in its entirety.

BACKGROUND OF THE INVENTION

15 This invention relates to methods and means for
rapid screening of target nucleic acid molecules for the
presence of sequence signatures. In preferred embodiments,
hybridization data is processed by a programmable digital
computer.

20 Polynucleotide arrays, such as the GeneChip® array
(Affymetrix, Inc., Santa Clara, CA, USA), can contain many
thousands of differently sequenced polynucleotide probes at
feature densities greater than five hundred thousand per 1
cm². Such arrays enable one to obtain nucleotide sequence
25 information from target nucleic acid molecules. The
information is obtained by performing a hybridization reaction
between the target nucleic acid molecule and the
polynucleotide probes on the polynucleotide array. The
location and identity of the probes to which the target has
30 hybridized, and the extent of hybridization, is determined.
Because hybridization between nucleic acids is a function of
their sequences, analysis of the sequence of the probes to
which the target has hybridized, as well as the extent of
hybridization, provides information about the sequence of the
35 target molecule.

Because polynucleotide arrays can have many
thousands of probes, hybridization reactions create large
amounts of raw data for analysis. Already, several ways of

processing such data have been developed. In one application, one examines hybridization between a target molecule and a set of probes that are based upon a reference nucleotide sequence. Probes in the set to which the target does not hybridize or hybridizes weakly indicate sequences in which the target differs from the reference sequence. Nucleic acid arrays have been used to interrogate single nucleotide differences between reference and target nucleic acid sequences. Examples include the identification of genetic variants of infectious diseases, such as HIV, or genetic diseases, such as cystic fibrosis.

Other ways of obtaining useful information from hybridization data would be of benefit to the scientific and medical communities.

SUMMARY OF THE INVENTION

The present invention involves a hierarchical method of array-based analysis in which single nucleotide base determination may or may not be one step. The present invention has several embodiments, many of which involve the determination of a sequence signature. Useful sequence signatures include polynucleotide or polypeptide sequence signatures, such as those defining protein domains, gene families, different genes in a genome, repeat sequences, or polymorphic forms of a gene. The methods involve performing hybridization assays between the target nucleic acid molecules to be screened and polynucleotide arrays designed to identify targets that contain the sequence signatures. The arrays contain probe sets. The probes in a set, taken together, have the sequence of the sequence signature, or variations upon that sequence. Thereby, the probes define the reference sequence signature and sequences related to the sequence signature. A hybridization assay between the target molecule and the probes in the array generates data about which probes the target has hybridized to, as well as the extent of hybridization, if that data is so desired. Computer programs are then used to process the data. By determining whether the target has hybridized to probes defining one or more reference sequences, or to probes defining sequences that deviate from

the reference sequences, one can determine whether the target has the same sequence or a sequence similar to one or more of the reference sequences. By selecting appropriate reference sequences to put on the array as probes, one can determine whether a target encodes a particular closely related polypeptide sequence signature, or is a member of a gene family, or has the sequence of a particular or closely related gene in the genome. One can also look at patterns of differences between target and reference sequences to identify novel gene families, new members of gene families, and the like. By identifying the similarities and/or differences between the reference and target sequences, one can also determine the position on the chromosome of a target nucleic acid molecule.

To determine whether a target nucleic acid molecule has a sequence signature, the following steps can be employed: providing an polynucleotide array comprising a set of polynucleotide probes that define the sequence signature; generating hybridization data by performing a hybridization reaction between the target nucleic acid molecule and the probes in the set and detecting hybridization between the target nucleic acid molecule and each of the probes in the set; and processing the hybridization data to determine whether the target nucleic acid molecule has the sequence signature. In certain embodiments, the sequence signature is a polypeptide sequence signature; the sequence signature contains variable positions or the step of processing is performed by a programmable digital computer. In another embodiment, if the sequence signature is an amino acid sequence signature, the array comprises sets of probes that define the degenerate set of nucleotide sequence signatures encoding the polypeptide sequence signature. In addition, or as an alternative to degenerate probe sets, useful probe sets can contain inosine, other generic bases, or mixtures of A, C, T, G at the 3d position of a codon site. Probe sets can also contain sequences that query the presence of polymorphic variants of a sequence signature.

One aspect of the invention provides a method of analyzing a nucleic acid sample, comprising selecting a hierarchy of assay techniques comprising at least a first and second assay. The first assay is selected to provide a determination of the presence or absence or variant of a first sequence signature and the second assay being selected to provide a determination of the presence or absence or variant of a second sequence signature. At least one of the assays employs a high-density nucleic acid array. One analyzes the nucleic acid sample using the first assay. The nucleic acid sample is optionally analyzed in a second assay depending upon the results of the first assay.

In a further embodiment, the first or second sequence signature is a conserved region of a gene family. In certain embodiments, the first or second sequence signature is a non-conserved region of a gene family. The method can additionally comprise determining the full length sequence of said nucleic acid sample.

The present invention also provides a method of selecting clones for analysis. This aspect of the invention provides a support having a variety of clones associated with it. The support is exposed to one or more polynucleotides under low, medium, or high stringency conditions to permit at least some hybridization between the clones and the polynucleotides. One identifies the clones that hybridize with the polynucleotides. Clones selected for analysis are those not identified as hybridizing to the clones.

In one embodiment of this method, the support is a high-density nucleic acid array.

Also provided is a method of narrowing a nucleic acid sample for analysis. The steps are providing a sample containing nucleic acids; analyzing whether the sample contains a sequence signature using a high-density nucleic acid array; and further analyzing the nucleic acid sample only if that sequence signature is not present.

This invention also provides a method for determining whether a target molecule has a sequence from a gene family member. The method involves providing an

polynucleotide array comprising, for each of at least two different gene family members, a set of polynucleotide probes that define a reference nucleotide sequence from the gene family member; generating hybridization data by performing a hybridization reaction between the target nucleic acid molecule and the probes in the sets and detecting hybridization between the target nucleic acid molecule and each of the probes in the sets; and processing the hybridization data to determine whether the target nucleic acid has the reference sequence from one of the gene family members. In one embodiment, the step of selecting the target nucleic acid molecule by determining whether the target hybridizes to a nucleic acid probe that hybridizes to a gene encoding the gene family members. In other embodiments, the step of processing is performed by a programmable digital computer; the polynucleotide array further comprises, for each of the gene family members, a probe set defining a highly conserved region of the gene and a probe set defining a highly variable region of the gene; the polynucleotide array further comprises, for each of the gene family members, probe sets defining at least two highly conserved regions of the gene and probe sets defining at least two highly variable regions of the gene; the region codes for an amino acid sequence and the array further comprises probe sets capable of defining the different nucleotide sequences encoding the amino acid sequence. In one embodiment, the method further comprises the step of determining the nucleotide sequence of the target nucleic acid molecule if the target does not have the sequence of the region of a gene family member.

In another aspect, the invention provides a computer program product for analyzing hybridization data comprising code that receives as input the sequence of an polynucleotide probe in each feature of an polynucleotide array; code that receives as input reference nucleotide sequences from a plurality of members of a gene family; code that identifies a set of features in the polynucleotide array having probes that define the nucleotide sequence sequences; code that receives as input hybridization data from a hybridization reaction

between a target nucleic acid molecule and polynucleotide probes in the polynucleotide array; code that processes the hybridization data to determine whether the target nucleic acid molecule has a sequence from any of the reference sequences; and a computer readable medium that stores the codes.

In another aspect this invention provides a method that involves determining whether a target nucleic acid molecule comprises a sequence from one of a set of genes. The method comprises providing a target nucleic acid molecule comprising nucleotide sequences from genomic DNA; providing an polynucleotide array comprising, for each gene in the set, polynucleotide probes that define at least one sequence signature from a unique region of the gene; generating hybridization data by performing a hybridization reaction between the target nucleic acid molecule and the probes in the sets and detecting hybridization between the target nucleic acid molecule and each of the probes in the sets; and processing the hybridization data to determine whether the target nucleic acid /comprises a sequence from the unique region of one of the genes. In one embodiment, the step of processing is performed by a programmable digital computer. In another embodiment, the unique region of the gene codes for an amino acid sequence. In a further embodiment, the polynucleotide array further comprises, for each of the unique regions, a set of polynucleotide probes whose sequences define the degenerate set of nucleotide sequences that encode the amino acid sequence. The probes in such embodiments can in addition or as an alternative comprise sequences that contain generic bases such as inosine particularly at the third codon position. As an even further additional or alternative, polynucleotide probes can have a mixture of A, C, T, and G in the third codon position within a single feature of a polynucleotide array.

In another aspect, this invention provides a computer program product for analyzing hybridization data comprising code that receives as input the sequence of an polynucleotide probe in each feature of an polynucleotide

array; code that receives as input sequence signatures from a unique region of a plurality of genes; code that identifies a set of features in the polynucleotide array having probes that define the sequence; code that receives as input hybridization data from a hybridization reaction between a target nucleic acid molecule and polynucleotide probes in the polynucleotide array; code that processes the hybridization data to determine whether the target nucleic acid molecule comprises a sequence from any of the sequence signatures; and a computer readable medium that stores the codes.

BRIEF DESCRIPTION OF THE FIGURES

Fig. 1A illustrates an example of a computer system used to execute software that can be used to analyze data generated by the present invention. Figure 1A shows a computer system 1 which includes a monitor 3, screen 5, cabinet 7, keyboard 9, and mouse 11. Mouse 11 may have one or more buttons such as mouse buttons 13. Cabinet 7 houses a CD-ROM drive 15 and a hard drive (not shown) that may be utilized to store and retrieve computer programs including code incorporating the present invention. Although a CD-ROM 17 is shown as the computer readable storage medium, other computer readable storage media including floppy disks, DRAM, hard drives, flash tape memory, tape, and the like may be utilized. Cabinet 7 also houses familiar computer components (not shown) such as a processor, memory, and the like.

Fig. 1B shows a system block diagram of computer system 1 used to execute software that can be used to analyze data generated by the present invention. As in Figure 1A, computer system 1 includes monitor 3 and keyboard 9. Computer system 1 further includes subsystems such as a central processor 102, system memory 104, I/O controller 106, display adapter 108, removable disk 112, fixed disk 116, network interface 118, and speaker 120. Removable disk 112 is representative of removable computer readable media like floppies, tape, CD-ROM, removable hard drive, flash memory, and the like. Fixed disk 116 is representative of an internal hard drive, DRAM, or the like. Other computer systems

suitable for use with the present invention may include additional or fewer subsystems. For example, another computer system could include more than one processor 102 (i.e., a multi-processor system) or memory cache.

5 Fig. 2 depicts tiling to detect signature sequence. The signature sequence, Asn-Gly-Lys-Ala-Met (SEQ ID NO:2), encompasses a degenerate set of 64 nucleotide sequences that encode it. One of these is ATTGGCAAAG CTATG (SEQ ID NO:1). A probe set of 6-mers based on single-increment tiling that defines this reference sequence is the set ATTGGC (1-6 of SEQ ID NO:1), TTGGCA (2-7 of SEQ ID NO:1), TGGCAA (3-8 of SEQ ID NO:1), GGCAAA (4-9 of SEQ ID NO:1), GCAAAG (5-10 of SEQ ID NO:1), CAAAGC (6-11 of SEQ ID NO:1), AAAGCT (7-12 of SEQ ID NO:1), AAGCTA (8-13 of SEQ ID NO:1), AGCTAT (9-14 of SEQ ID NO:1) and GCTATG (10-15 of SEQ ID NO:1). Another reference sequence within the degenerate is AACGGAAAGG CAATG (SEQ ID NO:___). A probe set of 6-mers based on single-increment tiling that defines this sequence is AACGGA (1-6 of SEQ ID NO:___), ACGGAA (2-7 of SEQ ID NO:___), CGGAAA (3-8 of SEQ ID NO:___), GGAAAG (4-9 of SEQ ID NO:___), GAAAGG (5-10 of SEQ ID NO:___), AAAGGC (6-11 of SEQ ID NO:___), AAGGCA (7-12 of SEQ ID NO:___), AGGCAA (8-13 of SEQ ID NO:___), GGCAAT (9-14 of SEQ ID NO:___), and GCAATG (10-15 of SEQ ID NO:___).

25 Fig. 3 depicts an example of a strategy for detecting members of a gene family. Four regions of three genes in a family are selected as signature sequences (in brackets). The nucleotide sequence signatures of each of genes 1, 2 and 3 are defined by probe sets that include probes A, B, C, D (variable region V₁); E, F, G, H (variable region V₂), I, J, K, L (constant region C₁) and M, N, O, P (constant region C₂)

35 Fig. 4 depicts an example of a strategy for detecting sequence signatures from a variety of genes. In this example, an polynucleotide array having 525 x 525 features is provided that contains probes with all possible 9-mer sequences. Two polypeptide signature sequences are checked, Asn-Gly-Lys-Ala-Met (SEQ ID NO:___) and Arg-Arg-Gly-Ser-Phe (a site recognized by protein kinase A). Particular

WO 98/12354

9

nucleotide reference sequences ATTGGCAAAG CTATG (SEQ ID NO:1) and CGCCGCGGAA GTTTT (SEQ ID NO:2) are defined by the probe sets a-g and h-n, respectively. The programmable digital computer contains code that identifies the location on the array of the features having the defining probes, indicated as letters on the array. In processing the hybridization information, it uses data from the features for the defining sets.

Fig. 5 depicts the nucleotide sequences of four regions of each of thirty-two members of the TGF- β family of genes. (SEQ ID NOS: 1-32.) The first region of thirty nucleotides is selected from a variable region just 5' of the mature region of the RXXR (SEQ ID NO:1) cleavage site. The second region of twenty-four nucleotides is a variable region selected from nucleotides encoding the first eight amino acids of the mature TGF- β polypeptide. The third region of twenty-seven nucleotides is selected from the region encoding the CXGXC (SEQ ID NO:2) sequence signature. The fourth region of eighteen nucleotides is selected from the region encoding the CXC conserved motif at the end of the mature coding region of TGF- β . The far right column indicates the clone name from which the sequences were selected.

Fig. 6 shows results of a hybridization assay between DNA from TGF- β clones hBMP-4 and hBMP-5 and an array tiled with the sequences described in Fig. 5 using single-increment and trellis tiling for all ninety-nine bases of each of the thirty-two TGF- β clones.

DETAILED DESCRIPTION OF THE INVENTION

I. DEFINITIONS

As used herein, the following terms have the following meanings.

The term "complementary" refers to the topological compatibility or matching together of interacting surfaces of a probe molecule and its target. Thus, the target and its probe can be described as complementary, and furthermore, the contact surface characteristics are complementary to each other. Complementary includes base complementary such as G is

complementary to C and A is the complement of T or U in the genetic code. Complementary also includes other forms of ligand-receptor (also known as ligand-anti-ligand) interactions, such as between other types of receptors and their agonists, antagonists, and other molecules that bind thereto or show some affinity therefor.

The term "probe" includes a surface-immobilized polynucleotide or other polymer that can be recognized by a particular target. Depending on context, the term "probe" refers both to individual polynucleotide molecules and to the collection of same-sequence polynucleotide molecules surface-immobilized at a discrete location. Probe and target are often used interchangeably depending upon the context; a probe can bind or become associated with a target as part of a ligand-anti-ligand pair. The probes and targets of the present invention can comprise the bases as found in nature or analogs thereof.

The term "target" refers to a molecule of interest. The probe is useful in obtaining information about the target: whether the target has an affinity for a given probe. Targets may be naturally-occurring or man-made nucleic acid molecules. Also, they can be employed in their unaltered state or as aggregates with other species. Targets may be associated, covalently or noncovalently, to a binding member, either directly or via a specific binding substance. Targets are sometimes referred to in the art as anti-probes. A "Probe-Target Pair" is formed when two macromolecules have combined through molecular recognition to form a complex.

A target molecule can hybridize to a probe set defining a reference whose sequence is at least partially complementary to that of the target. However, in order to make the comparison of target and reference sequences easier, the sequence of a target or, conversely, of a reference and the probes that interrogate it may be given herein as their complement.

The term "feature" refers to an area of a substrate having a collection of substantially same-sequence, surface-immobilized polynucleotide probes. Generally, one feature is

different than another feature if the probes of the different features have substantially different nucleotide sequences. Certain features, however, can be designed to have mixtures of sequences for example to query the 3 position of a codon sequence. In the context of light-directed polynucleotide synthesis, for example, a feature is a spatially-addressable synthesis site. See U.S. Patent Nos. 5,384,261; 5,143,854; 5,510,270; 5,593,139; 5,634,734; and WO/95/11995.

5 The term "polynucleotide array" refers to a substrate having polynucleotide probes with different, known sequences, at discrete, known locations associated with its surface. Polynucleotide arrays have at least two different features and a density of at least five hundred features per square cm. In certain embodiments the arrays can have a density of about 625, at least one thousand, at least 10 thousand, at least 100 thousand, at least one million or at least 10 million features of per square cm. The substrate can be, merely by way of example, silicon or glass and can have the thickness of a glass microscope slide or a glass cover slip. Substrates that are transparent to light are useful when the method of performing an assay on the chip involves optical detection. The term also refers to a probe array and the substrate to which it is attached that form part of a wafer.

10 "High-density polynucleotide arrays" contain at least 400 different polynucleotide sequences per cm².

15 To "screen" means analyze a sample for certain characteristics. One often screens samples to narrow the pool of material that is subjected to further analysis. The assay utilized to screen will be a function of the characteristic that one wants to identify. For example, one can screen using any of the following illustrative methods: immunological methods such as antibody assays, functional assays to see if a starting material exhibits a biological property such as enzyme activity or activity associated with certain cell types, hybridization assays, chemical assays, NMR, mass spectroscopy, chromatography, electron spin resonance, isoelectric focussing, electrophoresis, and the like. Useful

screening assays include combinations of these methods performed sequentially or at the same time. It will be apparent to those skilled in the arts that Northern, Southern, and Western blotting and various other methods can be employed. The term "pre-screen" means a screening step that takes place before another step. As used herein, the terms screen and prescreen can at times be used interchangeably.

The term "sequence" refers, depending on context, to the nucleotide (base) sequence of a nucleic acid or the amino acid sequence of a polypeptide.

The term "nucleic acid sequence signature" refers to a chosen or reference nucleotide sequence. Sequence signatures include nucleotide sequences at most 300, 250, 200, 150, 100, 75, 50, 30, 25 or at most 15 nucleotides in length. Sequence signatures include sequences less than 10, 15, 25, 30, 35, 45, 50, 60, 70, 80, 90, 100, 120, 135, 150, 175, 200, 250 and 300 nucleotides long. Sequence signatures also include any combination of these parameters. Nonlimiting examples of nucleic acid sequence signatures occurring in nature include, e.g., the Hogness Box, the TATA box, a homeobox, the CAAT box and Alu repeat sequences.

The term "polypeptide sequence signature" refers to an amino acid sequence. Amino acids in the sequence signature are selected from the group of twenty common amino acids and also include the less common amino acids. The collection of all amino acid sequences defined by the sequence signature is referred to as the polypeptide "signature set."

Polypeptide sequence signatures include amino acid sequences in which the identity of all, a majority or at least two of the amino acids (common amino acids, uncommon amino acids, or analogues thereof) is fixed. A sequence signature can be chosen to be fixed or variable. Sequence signatures also include amino acid sequences in which the set of amino acids that can occupy a variable position is selected from at most 15, at most 10, at most 5 or at most 2 of the twenty common amino acids. Other amino acids, including those known to those skilled in the biochemical arts as the less common amino acids, are also included. Polypeptide sequence

signatures also include amino acid sequences of being most 300, 250, 200, 175, 150, 100, at most 50, at most 10 or at most 5 amino acids in length. Polypeptide sequence signatures also include sequences less than 275, 225, 180, 140, 120, 100, 90, 800, 70, 50, 40, 30, 20, 10, or 5 amino acids in length.

Polypeptide sequence signatures also include any combination of these parameters. Examples of polypeptide sequence signatures include the zinc finger motif and other structural motifs including without limitation coils, loops, helices, turns, leucine zippers, symmetric dimer features, and combinations thereof; the consensus recognition sequence for protein kinase A, for beta globins, for immunoglobulins, for the TGF beta superfamily, for DNA binding proteins, steroid-hormone receptor superfamily, to name a few.

The term "gene" refers to a genomic nucleic acid sequence at a particular genetic locus whose exons encode a polypeptide.

The term "region" when referring to a gene, means a sub-sequence of at least 15 contiguous nucleotides within the nucleotide sequence of the gene or a sub-sequence of at least 5 amino acids within the amino acid sequence of a polypeptide encoded by a gene.

The term "unique region" refers to a sub-sequence that is not shared by two genes.

The term "gene family" refers to a collection of genes encoding at least one polypeptide domain whose amino acid sequences have at least 25% sequence identity over a comparison window of at least 20 amino acids. Such domains are related through common ancestry as a result of gene duplication and subsequent evolution. Many polypeptide domains are known in the art including, for example, the EGF domain, the immunoglobulin domain, the fibronectin type III domain, the cadherin-like domain, death effector domains (DED) to name a few. See, Vaux, D.L., *Cell*, Vol. 90, pp.389-390 (1997) and in its entirety; *Molecular Biology of the Cell*, 3rd Ed., Alberts et al., (1994). Protein domains are also discussed in more detail in R.F. Doolittle, *Annu. Rev. Biochem.*, (1995) 64:287-314.

Gene families frequently encode polypeptides sharing at least one highly conserved region. Two polypeptides share a "highly conserved region" if the polypeptides have a sequence identity of at least 60% over a comparison window of five amino acids, or if they share a sequence identity of at least 80 % over a comparison window of ten amino acids.

Polypeptide members encoded by a gene family, the protein family, can have highly variable regions. A "highly variable region" of a polypeptide encoded by a gene family member is a region of ten amino acids that has less than 50% sequence identity with the same region of a polypeptide of another gene family member. Protein families that can be interrogated using the present invention include the TNF family, the BCL-2 family, actins, the heat shock proteins, keratins, myosin, protein kinases, transcription factors, tubulins, egg shell proteins, alpha globin, beta-like globins, immunoglobins, ovalbumin, transplantation antigens, visual pigment protein, and vitellogenin as non-limiting examples. See, Vaux, D.L., *Cell*, Vol. 90, pp.389-390 (1997) and in its entirety; *Molecular Biology of the Cell*, 3rd Ed., Alberts et al., (1994); Avise, J.C., *Molecular Markers, Natural History and Evolution*, Chapman and Hall publishers (1994); Stryer, L., *Biochemistry*, 3rd. Ed. (1988); and Atassi, M.Z., *Molecular Immunology*, Marcel Dekker, Inc. (1984).

"Pseudogenes" are genomic regions that do not result in protein products in the organisms that contain them. Pseudogenes have sequence similarities to their true gene counterparts. Pseudogenes may arise from duplication of ancestral genes except that mutations contained in the pseudogene interfere with transcription or translation. Lodish et al., *Molecular Cell Biology*, 3rd. Ed, Scientific American, Inc., New York, New York (1995). As used herein, pseudogenes can be members of gene and protein families that contain their functional counterparts.

"Tandem repeat genes" or "tandemly repeated genes" encode identical or nearly identical proteins or functional RNAs. The copies can appear one after the other separated by spacer regions that can vary within an individual. Lodish et

al., *Molecular Cell Biology*, 3rd. Ed, Scientific American, Inc., New York, New York (1995). As used herein the tandem repeat genes are a subset of members of a gene family.

5 "Tandem repeats" unlike tandem repeat genes can be simply nucleic acids segments that are repeated but do not necessarily encode protein or functional RNAs. These too are subsets of gene families as that term is used herein.

10 "Simple-Sequence DNA" is largely composed of 5-10 base pair sequences repeated in long tandem repeats. Those skilled in the art will appreciate that simple-sequence DNA is often concentrated at certain regions of a chromosome. For example, a large amount of mouse simple sequence DNA is located near centromeres. Accordingly, such sequences can be used as markers and to localize specific chromosomal sites.

15 Lodish et al., *Molecular Cell Biology*, 3rd. Ed, Scientific American, Inc., New York, New York (1995).
"Intermediate repeat DNA" or moderately repeated DNA is dispersed throughout various genomes: those of mammals, *Drosophila*, and yeast for example. Certain intermediate repeat sequences are transposable DNA elements. Lodish et al.,

20 *Molecular Cell Biology*, 3rd. Ed, Scientific American, Inc., New York, New York (1995).

"Spacer DNA" has no currently known function other than to separate functional DNA sequences. Lodish et al.,

25 *Molecular Cell Biology*, 3rd. Ed, Scientific American, Inc., New York, New York (1995).

An polynucleotide probe has a sequence "selected from" a reference sequence if the sequence of the polynucleotide probe is a sub-sequence of the reference sequence. For example, the probe ATTGGC (1-6 of SEQ ID NO:1) has a sequence selected from the reference sequence ATTGGCAAAG CTATG (SEQ ID NO:1).

30 A set of polynucleotide probes "define" a reference sequence if the sequences of the polynucleotide probes are selected from the reference sequence and if, taken together, the sequences of the polynucleotide probes include the entire sequence of the reference sequence. A probe set also defines sub-sequences within it. For example, the probes ATTGGC (SEQ

ID NO:3) and GGCAAA (SEQ ID NO:4) also define the sub-sequence TTGGCA (SEQ ID NO:___).

5 The term "tiling strategy" refers to criteria used to select a probe set that defines a reference sequence or set of reference sequences.

"Block tiling" generally refers to a tiling strategy including a set of probes defining a reference sequence in which none of the probes in the set overlap in sequence. For example, the reference sequence ATTGGCAAAG CTATG (SEQ ID NO:1) can be blocked tiled by the set, ATTGG (1-5 of SEQ ID NO:1), CAAAG (6-10 of SEQ ID NO:1), and CTATG (11-15 of SEQ ID NO:1).

10 "Single-increment tiling" refers to a tiling strategy including a set of probes that defines a reference sequence in which each probe in the set overlaps in sequence with another probe in the set except for a terminal nucleotide. For example, the reference sequence ATTGGCAAAG CTATG (SEQ ID NO:1) can be single-increment tiled by the set ATTGGC (1-6 of SEQ ID NO:1), TTGGCA (2-7 of SEQ ID NO:1), TGGCAA (3-8 of SEQ ID NO:1), GGCAAA (4-9 of SEQ ID NO:1), GCAAAG (5-10 of SEQ ID NO:1), CAAAGC (6-11 of SEQ ID NO:1), AAAGCT (7-12 of SEQ ID NO:1), AAGCTA (8-13 of SEQ ID NO:1), AGCTAT (9-14 of SEQ ID NO:1) and GCTATG (10-15 of SEQ ID NO:1).

25 "Double-increment tiling" refers to a tiling strategy including a set of probes that defines a reference sequence in which each probe in the set overlaps in sequence with another probe in the set except for two, consecutive terminal nucleotides. For example, the reference sequence ATTGGCAAAG CTATG (SEQ ID NO:1) can be double-increment tiled by the set ATTGGC (1-6 of SEQ ID NO:1), TGGCAA (3-8 of SEQ ID NO:1), GCAAAG (5-10 of SEQ ID NO:1), AAAGCT (7-12 of SEQ ID NO:1), AGCTAT (9-14 of SEQ ID NO:1) and CTATG (11-15 of SEQ ID NO:1).

35 "Standard tiling" refers to a tiling strategy for a sub-sequence of a reference sequence. Standard tiling includes a set of probes as follows. All nucleotide positions in the sub-sequence are designated fixed, except for one, which is designated variable. One probe in the set has (or

complements) the sequence of the reference sub-sequence. The other probes in the set have the same nucleotide as the reference sub-sequence at the fixed position, but have a different nucleotide at the variable position, and no two probes in the set have the same sequence. Thus, for example, the reference sequence ATTGGCA (1-7 of SEQ ID NO:1) may be standard tiled with the set ATTGGCA (1-7 of SEQ ID NO:1), ATTA_tGCA (SEQ ID NO:___), ATT_tGCA (SEQ ID NO:___) and ATTcGCA (SEQ ID NO:___).

"Standard, single-increment tiling" refers to a tiling strategy that combines in a single-increment tiling strategy with a standard tiling strategy for each probe in the single-increment set. For example, the reference sequence ATTGGCAAA (1-9 of SEQ ID NO:1) can be tiled with a standard, single-increment strategy with the following probe sets:

- Reference: ATTGGCAAA (1-9 of SEQ ID NO:1)
 Probes: (1) ATTGGCA (1-7 of SEQ ID NO:1),
 ATTA_tGCA (SEQ ID NO:___),
 ATT_tGCA (SEQ ID NO:___),
 ATTcGCA (SEQ ID NO:___);
 (2) TTGGCAA (2-8 of SEQ ID NO:1),
 TTGaCAA (SEQ ID NO:___),
 TTG_tCAA (SEQ ID NO:___),
 TTGcCAA (SEQ ID NO:___); and
 (3) TGGCAAA (3-9 of SEQ ID NO:1),
 TGGaAAA (SEQ ID NO:___),
 TGG_tAAA (SEQ ID NO:___),
 TGGgAAA (SEQ ID NO:___).

Hybridization data derived from arrays including Standard, single-increment tiling for a reference sequence is useful in base-calling -- the determination that a base in a target is different than the base in the reference sequence, and the identification of that different base.

The tiling strategy can be carried out by placing probes in adjacent placing probes adjacent features on the array. For example, probes defining a sequence in single-increment tiling can be placed in a horizontal row in adjacent features on the array; and probe sets for standard tiling for each of the probes in the single-increment set can be placed in a vertical row beneath the reference sequence.

However, the power of a programmable digital computer can reduce or eliminate the need for customized

tiling. For example, a probe array can contain probes defining all possible polynucleotide 9-mers. The computer can carry in its memory the location of the feature containing the probe having any given 9-mer sequence. Then, relying on the reference sequence, the computer can identify the locations of all the probes that make up, for example, the single-increment tiling set for the reference sequence. Similarly, the computer can identify the location of all the probes making up the standard tiling set for each of the probes defining the reference sequence. Then, in processing hybridization data, the computer can be programmed to examine hybridization between target and probe at each of the feature locations defining the single-increment, standard tiling set.

The term "degenerate set" refers to the set of all nucleotide sequences that encode a particular polypeptide sequence signature.

The term "high discrimination hybridization conditions" refers to hybridization conditions in which a single base mismatch can be determined.

Stringency conditions useful in the practice of the present invention are set forth in Sandbrook et al., Molecular Cloning: A Laboratory Manual, 2d Ed. (1989).

"Base calling" refers to a process involving comparing the nucleotide sequence of a target molecule with a reference nucleotide sequence and identifying positions at which the nucleotide in the target molecule is different than the nucleotide in the reference sequence. "ID base calling" refers to the process of base calling further involving determining the identity of a nucleotide in the target molecule that is different than a nucleotide in the same position of the reference sequence.

A target nucleic acid sequence is of "unknown genetic origin" if it has not been identified to derive from a known genetic locus.

II. DESCRIPTION

Nucleic acid arrays have been used to interrogate single nucleotide differences between reference and target

nucleic acid sequences. The present invention, by contrast involves a hierarchical method of array-based analysis in which single nucleotide base determination may or may not be one step. For example, the present invention provides as a first step (or series of steps) the determination of whether a target nucleic acid contains a sequence signature. The sequence signature can comprise, for example, a set of repeat sequences, a conserved region among gene family members, or other multinucleotide groupings that are of interest. The possible outcomes are that a target contains the exact sequence signature in its entirety, lacks the sequence signature in its entirety, or contains one or more variations of the sequence signature. Those outcomes can be used for a variety of purposes: to set up a classification system for nucleic acids of interest based upon sequence signatures, to assign nucleic acids to known groupings of existing classification systems, to determine what if any further analysis is desired, to decide whether to retain or discard all or portions of a nucleic acid sample, and whether special storage or disposal methods are required.

One skilled in the art will appreciate that such outcomes are particularly relevant to diagnosis of patients and in the identification of disease. For example, one can first determine whether a sequence signature present in a possible pathogen or set of pathogens is present in nucleic acids obtained from a patient's body fluids or tissue as a first step in narrowing down a diagnosis. For example, one can determine whether a sequence signature of a mycobacterium is present. See PCT Application No. PCT/US92/02102 (published August 14, 1997, publication no. WO97/29212), is present. Using the methods taught in PCT/US92/02102, one can identify, for example, the presence of different mycobacterium species based on sequences signatures. Using the hierarchical methods of the present invention, if no mycobacterium is present, the presence of another suspected pathogen is explored.

The process can be repeated using different arrays having different sets of sequence signatures until the desired level of detail regarding the sequence of the target nucleic

acid is obtained. Alternatively, this multistep process can be carried out in a single experiment on one array having probes directed to multiple sequence signatures. The methods disclosed herein can also be employed using one or more arrays in serial or parallel fashion.

The present invention relies upon the outcomes of first determinations to make decisions or further determinations until the desired level of information is determined. The present invention also provides a method of probing for the presence or absence of sequence signatures and their variants in a binary or trinary fashion. Binary analysis asks whether a specific sequence signature is identified or not; a binary determination is a yes/no determination. A trinary analysis asks whether a specific sequence signature is present, absent, or whether a variant of that sequence signature is present; a trinary determination is a yes/no/variant determination. One will appreciate that a quaternary analysis can also ask whether a variant is absent, and so on. The hierarchies contemplated by the present invention include a first determination based upon binding of a target to a polymer array followed by at least one other array-based determination of interest.

In binary and trinary analyses in which the goal is novel gene discovery, often the most useful information is contained in those samples that do not contain a particular sequence (a no in both binary and trinary analysis) and those that contain a variant of a particular sequence (a variant in trinary analysis). When doing gene discovery, it is of enormous benefit to prescreen nucleic acids for those that contain a sequence that has already been identified (a yes in binary or trinary analysis). One can, for example, discard all of the samples that contain the known sequence signature to focus further study on only those nucleic acids that do not contain that sequence signature. When looking for new genes, much time, labor, and money is saved by narrowing the pool of samples for analysis using this embodiment of the present invention.

An example of the practice of the present invention involves determining whether a set of samples contains a nucleic acid that is a member of a gene family. As a first step or first series of steps, one can determine the level of

5 hybridization between nucleic acids contained in those samples and arrayed probes that interrogate the sequence of either strand of a DNA encoding at least a portion of a conserved region associated with a gene family of interest. The samples are then divided according to the results. For purposes of
10 illustration, assume that of a set of three patient samples, the nucleic acids derived from those samples showed the presence of the entire portion of the probed-for conserved sequence, the absence of the entire portion of the probed-for conserved sequence, and at least one variant of the portion of
15 the probed-for conserved sequence. The patient samples and/or nucleic acids derived therefrom are segregated or classified accordingly into new sets, 1, 2, and 3. If a greater level of detail regarding one or more of the segregated sets is
20 desired, further analysis is performed. The further analysis can differ for each of the 3 (in this case) sets.

Further inquiry of set 1 can include looking for any remaining portions of the conserved sequence that were not
interrogated in the first step(s). Based on the results of such further inquiry, set 1 can be further divided into three
25 subsets: A, B, and C. Those subsets can be analyzed to determine the presence of other conserved sequences or variants thereof. The subsets are further divided into three sub-subsets, for example. The sub-subset can be analyzed for
the presence of another sequence signature and so on.

30 Such an inquiry is particularly useful for identifying new members of gene families. For example, any samples or sets that contain nucleic acids containing sequence signatures that together identify those nucleic acids as
35 member of a gene family can be further screened using arrays that contain probes that query the non-conserved regions of the known gene family members. The arrays can be tiled to permit identification of sequence differences at the single

nucleotide level so that the novel sequence of the new family member is obtained.

It will also be appreciated that this method is also beneficial to industries involving large scale manufacture of polymers. In the biotechnical arts, for example, large scale recombinant protein synthesis can result in mixtures of recombinantly produced polypeptides. In certain cases, for example, *E. Coli* can insert so called "friendly" codons under certain fermentation conditions into some but not all of the polypeptide product. One can test recombinant protein for the presence of sequence signature variants as the first step or series of steps in a decision tree. That decision tree can involve the segregation of lots that contain the variants from those that do not. The variant lots can be discarded or further analyzed to the level of detail desired.

It will also be appreciated that the present invention involves the use of an array, preferably a high density array, in at least one of the steps of any method taught herein. The other steps may be performed using techniques known to those skilled in the arts. In addition, this application discusses the invention often in terms of nucleic acid arrays and nucleic acid analysis. Analysis of other materials and the use of other polymer arrays, including without limitation polypeptide and polysaccharide arrays is contemplated by the present invention. The hierarchy of analysis taught herein confers several advantages. One such advantage is conferred by looking, for example, at a small segment of sequence data - the sequence of the signature or sets of signatures instead of the full length sequence - to determine what if any further analysis is desired. By taking this hierarchical approach, the time, labor, cost, and amount of materials involved in handling and manipulating sample for analysis can be reduced.

The present invention not only provides this novel hierarchy of analysis, it further teaches that for discovery of previously uncharacterized molecules, the most useful information can be derived from analysis of those samples that have been shown to not contain, or segregated after a

screening step to decrease the likelihood of containing previously characterized sequence signatures or parts thereof. For example, a sample that contains a sequence signature from a conserved region of a gene family but does not contain the unique sequence signatures that characterize the different members of that gene family likely contains a novel member of the gene family. In such cases, further analysis of that sample is particularly desirable. For example, in such cases, and at that stage in the hierarchy of analysis, determination of the full length sequence or at least of the region that differs from the unique region of other members of the putative gene family is appropriate. Prior to or simultaneously with determination of the full length sequence at the single nucleotide level, one can obtain the footprint or bar code hybridization pattern. See WO97/29212 and EP Application No. 95307476.2, filed October 20, 1995, published as Ep 0717113A2, June 19, 1996. It will be appreciated that at least in certain circumstances, single nucleotide determination can be inferred from the footprint or other hybridization pattern.

The hierarchical methods of the present invention are particularly useful in the identification of gene family members; the discovery of new gene family members or other molecules; identify of nucleic acid fragments as being from or containing certain regions of a genome (human or otherwise); the handling or disposal of potentially hazardous materials including without limitation those containing prions, nucleic acid materials such as retroviruses; the segregation of materials into different biohazard groupings; epidemiological characterization and analyses; analysis of recombinantly or enzymatically manufactured biological materials such as nucleic acid (including without limitation antisense agents, ribozymes, promoter sequences, control sequences, restriction site sequences, capped sequences, tailed sequences, branched sequences, methylated sequences, vector sequences, analogues including peptide nucleic acids, and other sequences or varieties) protein or peptide (including without limitation insulin, growth factors, antibodies, endorphins, enkephalins,

and protein or peptide analogues) carbohydrate of all forms and analogues thereof, proteoglycans, and filamentous materials including without limitation those containing fibrins, actins, myosin, tropomyosin, troponin, and meromyosin; quality control and assurance for manufactured biological materials, natural or synthetic polymers, or other chemical materials; the narrowing of a clone pool and others.

All of the methods discussed herein can include correlating RNA levels with gene sequences of interest, the identification and use of expression patterns, and the narrowing of expression pattern information in a hierarchical fashion, or the selection, including by experimental design, of subsets of particular expression profiles. For example, one can look for the sequence signatures of enzymes involved in a particular metabolic pathway. If one or more of the sequence signatures are missing, a second assay can be for the sequence signatures of other enzymes that can or are thought to metabolize the excess accumulation of bioproducts that results from the enzyme deficiency screened for in the first assay.

A. Screening Methods

1. Analyzing For Sequence Signatures

In one aspect, this invention provides methods that involve analyzing a nucleic acid molecule for the presence of a sequence signature. Such analysis involves starting with an polynucleotide array that contains a set of probes that define the sequence signature; generating hybridization data by performing a hybridization assay between the target and the array and detecting hybridization between the target and the probes in the array, and processing the hybridization data to determine whether the target has the sequence signature.

The probes required on the polynucleotide array depend upon the sequence signature to be analyzed. The sequence signature can be, for example, an amino acid sequence or a nucleotide sequence. The sequence signature could define, for example, a polypeptide domain. The sequence signature could be a fixed sequence or a consensus sequence in

which certain of the positions are not fixed. Consider, for example, the consensus recognition sequence for protein kinase A: $RRX(S/T)Z$, wherein X is any amino acid and Z is a hydrophobic amino acid selected from valine, leucine or isoleucine (SEQ ID NO:3). The signature set for this amino acid sequence comprises $1 \times 1 \times 20 \times 2 \times 3$, or 240 different amino acid sequences. The number of nucleotide sequences that encodes this signature set equals the product of the number of codons that encode the possible amino acids at each position. In this case, a complete degenerate set for all the amino acid sequences in the signature set contains $6 \times 6 \times 61 \times (6 + 4) \times (4 + 6 + 3) = 285,480$ different reference nucleotide sequences. In contrast, the degenerate set of all possible nucleotide sequences encoding the single recognition sequence $RRGSV$ (SEQ ID NO:__), is $6 \times 6 \times 4 \times 6 \times 4 = 3456$ sequences. For recognition of certain sequence signatures, a generic array of all possible 8-mers, for examples, can be particularly useful. See PCT Application No. 97US/01603, filed January 22, 1997, PCT publication No. Wo97/17317.

In one embodiment, a polynucleotide array is selected that contains probe sets defining the complete degenerate set of nucleotide sequences encoding all possible amino acids in the amino acid signature sequence. However, alternative approaches are possible, depending upon the particular sort of analysis one wishes to undertake. For example, one might wish to analyze only a sub-set of the amino acid sequences in the polypeptide signature set. In this case, one can select an polynucleotide array whose probe sets define the degenerate sets encoding only those amino acid sequences. Or, one may wish to analyze for targets having selected codons encoding all or some of the amino acids in the polypeptide signature set. In this case, one can select an polynucleotide array having probe sets that define less than the degenerate set for any particular amino acid sequence, e.g., at least 25%, at least 50%, at least 75% or at least 95% of the degenerate set of sequences encoding the polypeptide sequence signature. (See Fig. 2.) Combinations of the two approaches are also possible. Useful probe sets can also

contain generic bases such as inosine or mixtures of A, C, T, G, or U at the equivalent of the third codon position in the sequence.

5 One then carries out a hybridization reaction in which the target nucleic acid sequence is contacted with the polynucleotide probe under hybridization conditions. If the target nucleic acid molecule is very long, one can optionally break the target into fragments and contact the array with the fragments. Usually the target or fragments thereof are
10 detectably labelled so that the positions at which they have hybridized can be determined.

After carrying out the hybridization reaction, hybridization is detected between selected probes and the target to generate hybridization data. This data usually
15 reflects the amount of hybridization, as determined by the strength of the detectable signal (fluorescence for example), between the target and the probes at a particular feature. One can use high, intermediate, or low discrimination hybridization conditions as desired.

20 The hybridization data is then processed, preferably by programmable digital computer, to determine whether the target contains a nucleotide reference sequence defined by any probe set. Processing the hybridization information can comprise determining the degree of fidelity of hybridization
25 between the target nucleic acid molecule and each probe in the set, whereby hybridization with high fidelity to all the probes in the set indicates that the target nucleic acid molecule has the sequence signature, and hybridization with high fidelity to a subset of the probes in the set indicates
30 that the target nucleic acid molecule has part of the sequence signature.

For example, suppose one desired to determine whether a target polynucleotide encoded the amino acid sequence RRGSV (SEQ ID NO:___). As stated above, 3456
35 nucleotide sequences encode this amino acid sequence. An array can be selected that includes probe sets using a single-increment tiling strategy defining the degenerate set of nucleotide sequences that encodes RRGSV (SEQ ID NO:___).

Suppose, further, that the target nucleic acid has the sequence CGACGAGGGTCTGTC (SEQ ID NO:___), which encodes RRGSV (SEQ ID NO:___). Under high discrimination hybridization conditions, this target sequence would hybridize to the single-increment probe set as depicted by asterisk:

5	Signature:	R R S V G	(SEQ ID NO:___)
	Reference:	CGACGAGGGTCTGTC	(SEQ ID NO:___)
	Probes:	*CGACGA	(SEQ ID NO:___),
		*GACGAG	(SEQ ID NO:___),
		*ACGAGG	(SEQ ID NO:___),
10		*CGAGGG	(SEQ ID NO:___),
		*GAGGGT	(SEQ ID NO:___),
		*AGGGTC	(SEQ ID NO:___),
		*GGGTCT	(SEQ ID NO:___),
		*GGTCTG	(SEQ ID NO:___),
15		*GTCTGT	(SEQ ID NO:___), and
		*TCTGTC	(SEQ ID NO:___).

20 By detecting hybridization between the target and each of the above probes, one can determine that the target molecule has the sequence of the reference sequence. One can also determine a hybridization pattern to identify the sequence signature.

25 An array that contained probe sets defining all degenerate sequences encoding RRGSV (SEQ ID NO:___) also would contain a probe set defining the sequence CGCCGAGGGTCCGGG (SEQ ID NO:___). It would hybridize to the target molecule as shown with an asterisk:

30	Signature:	R R S V G	(SEQ ID NO:___),
	Target:	CGACGAGGGTCTGTC	(SEQ ID NO:___),
	Reference:	CGCCGAGGGTCCGGG	(SEQ ID NO:___),
	Probes:	CGCCGA	(SEQ ID NO:___),
		GCCGAG	(SEQ ID NO:___),
		CCGAGG	(SEQ ID NO:___),
35		*CGAGGG	(SEQ ID NO:___),
		*GAGGGT	(SEQ ID NO:___),
		*AGGGTC	(SEQ ID NO:___),
		GGGTCC	(SEQ ID NO:___),
		GGTCCG	(SEQ ID NO:___),
40		GTCCGG	(SEQ ID NO:___), and
		TCCGGG	(SEQ ID NO:___).

45 Thus, upon detecting hybridization between the target and this probe set, one can determine that the target does not have this complete reference sequence, or is related to a sequence signature although it encodes the sequence RS.

The target may hybridize to part of a reference nucleotide sequence but it may not hybridize at positions representing particular codons. In this case, the target does not encode the polypeptide sequence signature, but may encode a related sequence signature which varies from the original as a result of a variable amino acid position. If the array contains probes defining sequence signatures that include such variable positions, the computer system can process the hybridization data from the probe sets defining these other sequence signatures, to determine whether the target encodes one of theses. If the target fails to hybridize to the probes defining a sequence signature, then the target does not encode this sequence signature.

The array need not include probes defining the degenerate set of nucleotide sequences encoding a polypeptide sequence signature. As an alternative to a degenerate set of nucleotide sequences, one can provide for generic bases such as inosine or mixtures of A, C, T, G, and U at what corresponds to the third codon position. In addition, one can employ footprint, molecular bar-coding, or other hybridization patterns to determine the presence of, absence of, or variance from the reference sequence signature.

In another embodiment of this method, the array further comprises probe sets selected for standard tiling of a reference sequence. Suppose, for example, that as a result of mutation, the target nucleic acid has the sequence CGA CGA tGG TCT GTC (SEQ ID NO:___), which encodes RRWSV (SEQ ID NO:___). A probe set that is standard tiled throughout the reference sequence may include probe sets that hybridize to the target as follows:

Signature:	R R S V G	(SEQ ID NO:___)
Reference:	CGACGAGGGTCTGTC	(SEQ ID NO:___)
Target:	CGACGAtGGTCTGTC	(SEQ ID NO:___)
Probes:	*CGACGA	(SEQ ID NO:___),
	CGtCGA	(SEQ ID NO:___),
	CGgCGA	(SEQ ID NO:___),
	CGcCGA	(SEQ ID NO:___),
	GACGAG	(SEQ ID NO:___),
	GAAgAG	(SEQ ID NO:___),
	GAtGAG	(SEQ ID NO:___),
	GAgGAG	(SEQ ID NO:___),
	ACGAGG	(SEQ ID NO:___),

5

10

15

20

25

30

ACaAGG	(SEQ ID NO:___)
ACtAGG	(SEQ ID NO:___)
ACcAGG	(SEQ ID NO:___)
CGAGGG	(SEQ ID NO:___)
CGtGGG	(SEQ ID NO:___)
CGgGGG	(SEQ ID NO:___)
CGcGGG	(SEQ ID NO:___)
GAGGGT	(SEQ ID NO:___)
GAaGGT	(SEQ ID NO:___)
*GAtGGT	(SEQ ID NO:___)
GACGGT	(SEQ ID NO:___)
AGGGTC	(SEQ ID NO:___)
AGaGTC	(SEQ ID NO:___)
AGtGTC	(SEQ ID NO:___)
AGcGTC	(SEQ ID NO:___)
GGGTCT	(SEQ ID NO:___)
GGaTCT	(SEQ ID NO:___)
GGtTCT	(SEQ ID NO:___)
GGcTCT	(SEQ ID NO:___)
*GGTCTG	(SEQ ID NO:___)
GGaCTG	(SEQ ID NO:___)
GGgCTG	(SEQ ID NO:___)
GGcCTG	(SEQ ID NO:___)
*GTCTGT	(SEQ ID NO:___)
GTaTGT	(SEQ ID NO:___)
GTtTGT	(SEQ ID NO:___)
GTgTGT	(SEQ ID NO:___)
*TCTGTC	(SEQ ID NO:___)
TCaGTC	(SEQ ID NO:___)
TCgGTC	(SEQ ID NO:___)
TCcGTC	(SEQ ID NO:___)

From this information, one can determine that the target does not encode the signature sequence, but has the sequence:

35	*CGACGA	(SEQ ID NO:___)
	*GAtGGT	(SEQ ID NO:___)
	*GGTCTG	(SEQ ID NO:___)
	*GTCTGT	(SEQ ID NO:___)
	*TCTGTC	(SEQ ID NO:___) or
	CGACGAtGGTCTGTC	(SEQ ID NO:___), which

40

encodes

R R W S V.

45

Software such as the GeneChip® software from Affymetrix, Inc. (Santa Clara, CA, USA) can be used to analyze the hybridization data. See also International Publication No. WO 97/28212, European Patent Application Publication No. EP 0717113A2 (European Patent Application No. 95307476.2).

50

2. Screening For Members Of A Gene Family

In another aspect, this invention provides methods for determining whether a target nucleic acid molecule encodes a member of a gene family. This method is useful for
5 determining whether a target molecule is a known member of a family, or a new, previously unknown, member. In selecting arrays for this type of screening, several parameters can be varied.

One parameter is the number of gene family members
10 whose sequences are used on the array. Probe sets defining sequences from at least one and more preferably at least two members of the family are used on the array. However, for the identification of new family members, one preferably creates arrays containing probe sets defining sequences from all known
15 members of the family.

Another parameter that can be varied is the number of sequence signatures from each member of the gene family that are defined by probe sets on the array. A comparison of the amino acid and nucleotide sequences of known members of a
20 gene family reveals both highly conserved and variable sequence regions. Conserved regions, because they share a higher degree of identity between members, are more useful for determining whether a target encodes a member of the family. Variable regions, because they are the most distinct, are more
25 useful for discriminating between members of the family and for indicating whether a target encodes a new member of the family. Accordingly, arrays used for screening members of a family contain probe sets defining at least one sequence signature from each member of the gene family.

Another parameter that can be varied, related to the
30 second parameter, is the number of nucleotide sequences within a degenerate set encoding an amino acid signature sequence from one or more of the gene family members from which probe sets are chosen. For example, a nucleic acid signature
35 sequence from a member of a gene family, if it is within the coding region of the gene, encodes an amino acid sequence. Probe sets can be selected that define not only the reference nucleotide sequence, but members of the degenerate set that

5 encode the same amino acid sequence as the reference
nucleotide sequence. Such probe sets are useful in
identifying polymorphisms of any gene family member, as well
as new members of the family. Generic bases and probes having
mixtures of bases at certain codon positions, such as the
third codon position can also be employed.

10 Another parameter is the length of the sequence
signature. While there is no particular size limit,
generally, sequence signatures preferably at least 15
nucleotides long. A collection of sequence signatures
totalling between 75 and 125 nucleotides spread among about 4
signatures is particularly useful.

15 Any nucleic acid molecule can be used as a target
molecule in this method. However, often, the target is a
molecule that has been pre-screened in accordance with the
teachings of the present invention so that there is reason to
believe the target may be a member of the gene family. For
example, one may screen a DNA library with probes (which can
include degenerate sets, generic bases, and mixtures of
20 nucleotides at certain positions) having a sequence selected
from one or more members of the gene family. Depending upon
the stringency of the hybridization conditions used, the probe
may hybridize to sequences more closely or more distantly
related to the probe. Thus, the target sequence can be one
25 that hybridizes under a selected set of hybridization
conditions to a probe having the reference sequence.

The hybridization data generated from a
hybridization reaction between the target and the probes on
the array is processed to determine whether the data is
30 consistent with the target nucleic acid being a member of the
gene family. This can involve, for example, base calling the
target sequence over at least a sequence signature for a
conserved region of the gene or the determination of whether
the overall pattern expected for that sequence signature is
35 present.

The hybridization data may indicate that the target
molecule has sequences that are identical to that of a known
member of the gene family. However, if the hybridization data

indicates that there are differences between the target sequence and the reference sequences, the extent of the differences provides further information about the identity of the target sequence.

5 For example, if the differences are few enough, their location and identity can in certain embodiments be determined by ID base calling using, e.g., arrays that employ single increment, standard tiling. In this case, the information is consistent with the target being one of the
10 known gene family members, possibly including allelic forms of the gene.

If there are significant differences between the target and the probe sets, then the hybridization is generally quite weak in the regions that differ. In this case, the
15 target is identified as containing an insert that is not a previously known member of the family. The practitioner then can decide whether the clone is worth sequencing to determine if it is actually a member of the family, and, if so, how it differs from the other members.

20 3. Screening For Gene Sequences

As we move into a world in which all the genes of the human and other genomes are identified and sequenced, the focus of much nucleic acid analyses will be the identification
25 of which genes are present in a particular sample. Such identification is particularly useful in the hierarchical methods of the present invention. Accordingly, this invention also provides methods of determining whether a target nucleic acid molecule has a nucleotide sequence from any of a set of
30 genes. The methods involve providing an array with probe sets defining sequence signatures from the gene set. Hybridization data is collected from a hybridization reaction between the target and the probes on the array. The data is analyzed to determine whether the target contains the sequence signature
35 of from one of the genes in the set.

The hybridization data can be processed in the following manner. The extent of hybridization between the probes that define each sequence signature and the target can

be determined. If the target has a sequence closely related to one of the sequence signatures, the degree of hybridization between the target and the probe set that define the sequence signature of that gene will be strong compared to the hybridization signal with other sequence signatures defined in the array. This assists in identifying sequence signatures in the target. Computerized methods for analyzing hybridization data from nucleic acid arrays are taught by WO97/29212; EP publication No. 95307476.2.

Preferably the sequence signatures are unique to the genes in the set. A sequence signature of about 20 nucleotides suffices in most cases to uniquely identify a gene. The signature sequence can be from a transcribed or coding region of the gene. Such arrays are useful for example for determining the identity of target cDNA molecules, variants of genomic clones.

One can obtain further information about a target sequence by providing arrays with probe sets in single-increment, standard tiling systems for each nucleotide in each of the sequence signatures. In this case, when the target does not perfectly match a sequence signature, differences between the two sequences can be determined by base calling. Alternatively, recognition of the hybridization pattern is employed.

B. Performing Hybridization Assays

Hybridization assays on substrate-bound polynucleotide arrays involve a hybridization step and a detection step. In the hybridization step, a hybridization mixture containing the target and, preferably, a hybridization optimizing agent, such as an isostabilizing agent, denaturing agent or renaturation accelerant, is brought into contact with the probes of the array and incubated at a temperature and for a time appropriate to allow hybridization between the target and any complementary probes. Usually, unbound target molecules are then removed from the array by washing with a wash mixture that does not contain the target, such as hybridization buffer. This leaves only bound target

molecules. In the detection step, the probes to which the target has hybridized are identified. Since the nucleotide sequence of the probes at each feature is known, identifying the locations at which target has bound provides information about the particular sequences of these probes.

The hybridization mixture includes the target nucleic acid molecule and a hybridization optimizing agent in an appropriate solution, i.e., a hybridization buffer. The target nucleic acid molecule is present in the mixture at a concentration between about 0.005 nM and about 50 nM preferably between about 0.5 nM and 5 nM or, more preferably, about 1 nM and 2 nM. The target nucleic acid molecule preferably includes a detectable label, such as a fluorescent label.

Betaines and lower tetraalkyl ammonium salts are examples of isostabilizing agents. Denaturing agents are compositions that lower the melting temperature of double stranded nucleic acid molecules by interfering with hydrogen bonding between bases in a double-stranded nucleic acid or the hydration of nucleic acid molecules. Denaturing agents include formamide, formaldehyde, DMSO ("dimethylsulfoxide"), tetraethyl acetate, urea, GuSCN, glycerol and chaotropic salts. Hybridization accelerants include heterogenous nuclear ribonucleoprotein ("hnRP") A1 and cationic detergents such as, preferably, CTAB ("cetyltrimethylammonium bromide") and DTAB ("dodecyl trimethylammonium bromide"), and, also, polylysine, spermine, spermidine, single stranded binding protein ("SSB"), phage T4 gene 32 protein and a mixture of ammonium acetate and ethanol.

The hybridization mixture is placed in contact with the array and incubated. Contact can take place in any suitable container, for example, a dish or a cell specially designed to hold the array and to allow introduction of the fluid into and removal of it from the cell so as to contact the array. Generally, incubation will be at temperatures normally used for hybridization of nucleic acids, for example, between about 20° C and about 75° C, e.g., about 25° C, about 30° C, about 35° C, about 40° C, about 45° C, about 50° C,

about 55° C, about 60° C or about 65° C. For probes longer than about 14 nucleotides, 20° C - 50° C is preferred. For shorter probes, lower temperatures are preferred. The target is incubated with the probe array for a time sufficient to allow the desired level of hybridization between the target and any complementary probes in the array. Using a hybridization temperature of 25° C can yield a very clear signal, usually in at least 30 minutes to two hours, but it may be desirable to hybridize longer, i.e., about 15 hours.

After incubation with the hybridization mixture, the array usually is washed with the hybridization buffer, which also can include the hybridization optimizing agent. These agents can be included in the same range of amounts as for the hybridization step, or they can be eliminated altogether. Then the array can be examined to identify the probes to which the target has hybridized.

C. Preparation of Target Samples

The target polynucleotide whose sequence is to be determined can be isolated from a clone, a cDNA, genomic DNA, RNA, cultured cells, or a tissue sample. If the target is genomic, the sample may be from any tissue (except exclusively red blood cells). For example, whole blood, peripheral blood lymphocytes or PBMC, skin, hair or semen are convenient sources of clinical samples. These sources are also suitable if the target is RNA. Blood and other body fluids are also a convenient source for isolating viral nucleic acids. If the target is mRNA, the sample is obtained from a tissue in which the mRNA is expressed. If the polynucleotide in the sample is RNA, it is usually reverse transcribed to DNA. DNA samples or cDNA resulting from reverse transcription are usually amplified, e.g., by PCR. Depending on the selection of primers and amplifying enzyme(s), the amplification product can be RNA or DNA. Paired primers are selected to flank the borders of a target polynucleotide of interest. More than one target can be simultaneously amplified by multiplex PCR in which multiple paired primers are employed. If the target is a long polynucleotide, it may be appropriate to fragment the

target into smaller pieces before performing the hybridization reaction. As used herein, the detection of hybridization between a target and probes on an array includes performing the hybridization reaction with all or portions of the target.

5 The target can be labelled at one or more nucleotides during or after amplification. For some target polynucleotides (depending on size of sample), e.g., episomal DNA, sufficient DNA is present in the tissue sample to dispense with the amplification step. Preferred labels
10 include fluorescent labels, chemi-luminescent labels, bioluminescent labels, and colorimetric labels, among others. Most preferably, the label is a fluorescent label such as a fluorescein, a rhodamine, a polymethine dye derivative, a phosphor, and so forth. Commercially available fluorescent
15 labels include, *inter alia*, fluorescein phosphoramidites such as Fluoreprime (Pharmacia, Piscataway, NJ), Fluoredate (Millipore, Bedford, MA) and FAM (ABI, Foster City, CA).

Useful light scattering labels include large colloids, and especially the metal colloids such as those from
20 gold, selenium, silver, tin, and titanium oxide.

Radioactive labels include, for example, ^{32}P . This label can be detected by a phosphorimager. Detection, of course, depends on the resolution of the imager. Phosphorimagers are available having resolution of 50 microns.
25 Accordingly, this label is currently useful with chips having features of at least that size.

In one embodiment, biotinylated bases are incorporated into the target nucleic acid. Hybridization is detected by staining with streptavidin-phycoerythrin.

30 When the target strand is prepared in single-stranded form as in preparation of target RNA, the sense of the strand should of course be complementary to that of the probes on the chip. This is achieved as an example by appropriate selection of primers used for any amplification of
35 the target. Also, the array can contain probes for both strands.

The target is preferably fragmented before application to the chip to reduce or eliminate the formation

of secondary structures in the target and reduce any overhang interaction. The average size of targets segments following hybridization is usually larger than the size of probe on the chip.

5

D. Substrate-Associated Polynucleotide Arrays

Substrate-associated polynucleotide arrays used in the assays of this invention typically include between about 5×10^2 and about 10^8 features per square centimeter, or between
10 about 10^4 and about 10^7 , or between about 10^5 and 10^6 .

Preferably, the arrays are produced through spatially directed polynucleotide synthesis. As used herein, "spatially directed polynucleotide synthesis" refers to any method of directing the synthesis of an polynucleotide to a
15 specific location on a substrate. Methods for spatially directed polynucleotide synthesis include, without limitation, light-directed polynucleotide synthesis, microlithography, application by ink jet, microchannel deposition to specific locations and sequestration with physical barriers. In
20 general these methods involve generating active sites, usually by removing protective groups; and coupling to the active site a nucleotide which, itself, optionally has a protected active site if further nucleotide coupling is desired.

In one embodiment substrate-bound polynucleotide
25 arrays are synthesized at specific locations by light-directed polynucleotide synthesis. The pioneering techniques of this method are disclosed in U.S. Patent No. 5,143,854; PCT WO 92/10092; PCT WO 90/15070; and United States Application Serial Nos. 08/249,188, filed May 24, 1994, 07/624,120, filed
30 December 6, 1990, and 08/082,937, filed June 25, 1993. In a basic strategy of this process, the surface of a solid support modified with linkers and photolabile protecting groups is illuminated through a photolithographic mask, yielding reactive hydroxyl groups in the illuminated regions. A 3'-O-phosphoramidite-activated deoxynucleoside (protected at the
35 5'-hydroxyl with a photolabile group) is then presented to the surface and coupling occurs at sites that were exposed to light. Following the optional capping of unreacted active

5 sites and oxidation, the substrate is rinsed and the surface is illuminated through a second mask, to expose additional hydroxyl groups for coupling to the linker. A second 5'-protected, 3'-O-phosphoramidite-activated deoxynucleoside (C-X) is presented to the surface. The selective photodeprotection and coupling cycles are repeated until the desired set of products is obtained. Photolabile groups are then optionally removed and the sequence is, thereafter, optionally capped. Side chain protective groups, if present, are also removed. Since photolithography is used, the process can be miniaturized to generate high-density arrays of polynucleotide probes.

10 This general process can be modified. For example, the nucleotides can be natural nucleotides, chemically modified nucleotides or nucleotide analogs, as long as they have activated hydroxyl groups compatible with the linking chemistry. The protective groups can, themselves, be photolabile. Alternatively, the protective groups can be labile under certain chemical conditions, e.g., acid. In this example, the surface of the solid support can contain a composition that generates acids upon exposure to light. Thus, exposure of a region of the substrate to light generates acids in that region that remove the protective groups in the exposed region. Also, the synthesis method can use 3'-protected 5'-O-phosphoramidite-activated deoxynucleoside. In this case, the polynucleotide is synthesized in the 5' to 3' direction, which results in a free 5' end.

20 The general process of removing protective groups by exposure to light, coupling nucleotides (optionally competent for further coupling) to the exposed active sites, and optionally capping unreacted sites is referred to herein as "light-directed nucleotide coupling."

25 Tiling strategies for creating probe arrays adapted for various tasks, such as re-sequencing are described in U.S. patent application serial no. 08/510,521, filed August 2, 1995 and International application PCT/US94/12305, filed October 26, 1994.

If desired, the substrate-bound polynucleotide array can be appropriately packaged for use in a chip reader. One such apparatus is disclosed in International Publication No. WO 95/33846.

5 Probes may be laid out on an polynucleotide array with a specifically defined positional relationship. For example, the probes in the set can be positioned in adjacent features on the array. However, hybridization data from an polynucleotide array normally will be processed by a
10 programmable digital computer. The computer memory can be programmed to remember the sequence of each probe at each feature on the array. Consequently, one may provide an polynucleotide array or set polynucleotide arrays containing all possible sequences of probes of a given length. For
15 example, a chip having 525 by 525, or 275,625, features can contain all nine-mer probes having all possible nucleotide sequences of 9 nucleotides ($4^9 = 262,144$). Using any selected tiling strategy, the programmable computer can identify the set of features containing probes that define any given
20 reference sequence. Then, the computer can be programmed to process hybridization data from the probe set that defines a reference sequence.

E. Detecting Fluorescently Labelled Probes

25 Determining a signal generated from a detectable label on an array requires an polynucleotide array or chip reader. The nature of the polynucleotide array reader depends upon the particular type of label attached to the target molecules.

30 In one embodiment the chip reader comprises a body for immobilizing the polynucleotide array. Excitation radiation, from an excitation source having a first wavelength, passes through excitation optics from below the array. The excitation optics cause the excitation radiation to excite a region of an polynucleotide array on the
35 substrate. In response, labeled material on the sample emits radiation which has a wavelength that is different from the excitation wavelength. Collection optics, also below the

array, then collect the emission from the sample and image it onto a detector. The detector generates a signal proportional to the amount of radiation sensed thereon. The signals can be assembled to represent an image associated with the plurality of regions from which the emission originated.

According to one embodiment, a multi-axis translation stage moves the polynucleotide array in order to position different areas to be scanned, and to allow different locations of an array to be interrogated. As a result, a 2-dimensional image of the polynucleotide array is obtained.

The polynucleotide array reader can include an auto-focusing feature to maintain the sample in the focal plane of the excitation light throughout the scanning process. Further, a temperature controller may be employed to maintain the sample at a specific temperature while it is being scanned. The multi-axis translation stage, temperature controller, auto-focusing feature, and electronics associated with imaging and data collection are managed by an appropriately programmed digital computer.

In one embodiment, a beam is focused onto a spot of about 2 μm in diameter on the surface of the array using, for example, the objective lens of a microscope or other optical means to control beam diameter. (See, e.g., United States patent application 08/195,889, filed February 10, 1994).

In another embodiment, fluorescent probes are employed in combination with CCD imaging systems. Details of this method are described in United States Application Serial Number 08/301,051, filed September 2, 1994. In many commercially available microplate readers, typically the light source is placed above an array, and a photodiode detector is below the array. For the present methods, the light source can be replaced with a higher power lamp or laser. In one embodiment, the standard absorption geometry is used, but the photodiode detector is replaced with a CCD camera and imaging optics to allow rapid imaging of the array. A series of Raman holographic or notch filters can be used in the optical path to eliminate the excitation light while allowing the emission to pass to the detector. In a variation of this method, a

fiber optic imaging bundle is utilized to bring the light to the CCD detector. In another embodiment, the laser is placed below the polynucleotide array and light directed through the transparent wafer or base that forms the bottom of the polynucleotide array. In another embodiment, the CCD array is built into the wafer of the polynucleotide array.

The choice of the CCD array will depend on the number of polynucleotides in each array. If 2500 features of sequence-specific polynucleotides nominally arranged in a square (50 x 50) are examined, and 6 lines in each feature are sampled to obtain a good image, then a CCD array of 300 x 300 pixels is desirable in this area. However, if an individual array has 48,400 features (220 x 220) then a CCD array with 1320 x 1320 pixels is desirable. CCD detectors are commercially available from, e.g., Princeton Instruments, which can meet either of these requirements.

The detection device also can include a line scanner, as described in United States patent application 08/301,051, filed September 2, 1994. Excitation optics focuses excitation light to a line at a sample, simultaneously scanning or imaging a strip of the sample. Surface-bound fluorescent labels from the array fluoresce in response to the light. Collection optics image the emission onto a linear array of light detectors. By employing confocal techniques, substantially only emission from the light's focal plane is imaged. Once a strip has been scanned, the data representing the 1-dimensional image are stored in the memory of a computer. According to one embodiment, a multi-axis translation stage moves the device at a constant velocity to continuously integrate and process data. Alternatively, galvometric scanners or rotating polyhedral mirrors may be employed to scan the excitation light across the sample. As a result, a 2-dimensional image of the sample is obtained.

In another embodiment, collection optics direct the emission to a spectrograph which images an emission spectrum onto a 2-dimensional array of light detectors. By using a spectrograph, a full spectrally resolved image of the array is obtained.

The read time for an polynucleotide array will depend on the photophysics of the fluorophore (i.e., fluorescence quantum yield and photodestruction yield) as well as the sensitivity of the detector. For fluorescein, sufficient signal-to-noise to read a chip image with a CCD detector can be obtained in about 30 seconds using 3 mW/cm² and 488 nm excitation from an Ar ion laser or lamp. By increasing the laser power, and switching to dyes such as CY3 or CY5 which have lower photodestruction yields and whose emission more closely matches the sensitivity maximum of the CCD detector, one easily is able to read each array in less than 5 seconds.

F. Data Analysis

Data generated in hybridization assays is most easily analyzed with the use of a programmable digital computer. The computer program generally contains a readable medium that stores the codes. Certain files are devoted to memory that includes the location of each feature and the sequence of the polynucleotide probe at that feature. Because analysis often involves comparing the sequence of a target to a reference sequence, the program also can include in its memory the reference sequence. Using this information, the program can then identify the set of features on the array whose probes define the reference sequence in the selected tiling strategy. The computer also contains code that receives as input hybridization data from a hybridization reaction between a target nucleic acid molecule and polynucleotide probes in the polynucleotide array. The computer also contains code that processes the hybridization data. The computer program also can include code that receives instructions from a programmer as input.

The computer can transform the data into another format for presentation. Data analysis can include the steps of determining, e.g., fluorescent intensity as a function of substrate position from the data collected, removing "outliers" (data deviating from a predetermined statistical distribution), and calculating the relative binding affinity

of the targets from the remaining data. The resulting data can be displayed as an image with color in each region varying according to the light emission or binding affinity between targets and probes therein.

5 One application of this system when coupled with the CCD imaging system that speeds performance when the detection step involves hybridization of a labeled target polynucleotide with an polynucleotide in the array is to obtain results of the assay by examining the on- or off-rates of the
10 hybridization. In one version of this method, the amount of binding at each address is determined at several time points after the targets are contacted with the array. The amount of total hybridization can be determined as a function of the kinetics of binding based on the amount of binding at each
15 time point. Thus, it is not necessary to wait for equilibrium to be reached. The dependence of the hybridization rate for different polynucleotides on temperature, sample agitation, washing conditions (e.g., pH, solvent characteristics, temperature) can easily be determined in order to maximize the
20 conditions for rate and signal-to-noise. Alternative methods are described in Fodor et al., United States patent 5,324,633, incorporated herein by reference.

 The dependence of the hybridization rate for different polynucleotides on temperature, sample agitation,
25 washing conditions (e.g., pH, solvent characteristics, temperature) can easily be determined in order to maximize the conditions for rate and signal-to-noise.

G. Mechanics of Assays

30 Assays on polynucleotide arrays generally include contacting an polynucleotide array with a labelled sample under the selected reaction conditions, optionally washing the array to remove unreacted molecules, and analyzing the biological array for evidence of reaction between target
35 molecules the probes. These steps involve handling fluids. These steps can be automated using automated fluid handling systems for concurrently performing the detection steps on the array. Fluid handling allows uniform treatment of samples in

the wells. Microtiter robotic and fluid-handling devices are available commercially, for example, from Tecan AG.

The chip can be manipulated by a fluid-handling device. This robotic device can be programmed to set appropriate reaction conditions, such as temperature, add reagents to the chip, incubate the chip for an appropriate time, remove unreacted material, wash the chip substrate, add reaction substrates as appropriate and perform detection assays. The particulars of the reaction conditions are chosen depends upon the purpose of the assay, for example hybridization of a probe or attachment of a label to polynucleotides.

If desired, the chip can be appropriately packaged for use in chip reader. One such apparatus is disclosed in United States patent application 08/255,682, filed June 8, 1994.

H. Substrate-associated Polynucleotide Array Manufacture

In making a chip, the substrate and its surface preferably form a rigid support on which the sample can be formed. The substrate and its surface are also chosen to provide appropriate light-absorbing characteristics. For instance, the substrate may be functionalized glass, Si, Ge, GaAs, GaP, SiO₂, SiN₄, modified silicon, or any one of a wide variety of gels or polymers such as (poly)tetrafluoroethylene, (poly)vinylidenedifluoride, polystyrene, polycarbonate, or combinations thereof. Other substrate materials will be readily apparent to those skilled in the art upon review of this disclosure. In a preferred embodiment the substrate is flat glass or silica.

Surfaces on the solid substrate usually, though not always, are composed of the same material as the substrate. Thus, the surface may be composed of any of a wide variety of materials, for example, polymers, plastics, resins, polysaccharides, silica or silica-based materials, carbon, metals, inorganic glasses, membranes, or any of the above-listed substrate materials. In one embodiment, the surface

will be optically transparent and will have surface Si-OH functionalities, such as those found on silica surfaces.

Preferably, polynucleotides are arrayed on a chip in addressable rows and columns. Technologies already have been developed to read information from such arrays. The amount of information that can be stored on each chip depends on the lithographic density which is used to synthesize the wafer. For example, if each feature size is about 100 microns on a side, each chip can have about 10,000 probe addresses (features) in a 1 cm² area.

The following examples are offered by way of illustration, not by way of limitation.

EXAMPLE

The method of the invention was used to screen for new members of the TGF- β superfamily of proteins. There are currently 32 known members of this family. Clone libraries were created from genomic material based on hybridization to nucleic acid probes in solution that contain sequences complementary to sequence motifs that are indicative of members of this gene family. The genomic inserts were approximately 15 kb in size. Most of the inserts contain sequences from previously known members of the family.

Conventional approaches involve sequencing these 15 kb inserts over and over, most of the time only to find that the insert contains a family member that has already been identified. The method of this invention replaced those laborious and time consuming steps with a faster, easier screening method that can identify which clones contain known members of the family, and which few clones out of the large library are worth investigating in greater detail.

TGF- β clone screening polynucleotide array:

The array contained over 12,000 features with different probes with single-increment, 4-base trellis tilings for 99 bases for each of the 32 known members of the TGF- β family (see Fig. 5). The 99 bases were from 4 different regions of the genes and the contiguous regions range in size

from 18 to 30 bases. The interrogated regions were chosen based on a few criteria: they include regions that are (a) reasonably well conserved (highly conserved at the amino acid level, but less so at the DNA level) and that serve as
5 identifiers of the protein family, (b) highly variable and serve as unique identifiers of individual members of the family, and (c) not near expected intron/exon boundaries.

TGF- β clone samples for hybridization:

10 Either DNA or RNA can be produced from a clone using standard methods, e.g., nucleic acid extractions followed by PCR or in vitro transcription, with labeled bases incorporated during the polymerization step. Fragmented single-stranded
15 DNA or RNA can be used in the hybridization as well as fragmented double stranded DNA. The hybridizations are done in either 6XSSPE-T or 3M TMACl-T (buffered with Tris to avoid having any Na ions in the hybridization solution), and generally at temperatures above 30°C to improve discrimination and to reduce cross-hybridization (this is more important in
20 this application than for some re-sequencing applications because the samples include ~15,000 bases). If labeled RNA is used, samples are fragmented with heat in the presence of Mg²⁺. If DNA is used, samples are fragmented by treatment with DNase I prior to hybridization. This works with both
25 double stranded DNA or with DNA that is made single-stranded following PCR by degradation of one of the strands using lambda exonuclease.

Examples and data analysis:

30 Following hybridization and reading of the arrays, the images are analyzed using the TGF report GeneChip software (Affymetrix, Inc., Santa Clara, CA, USA). Base calls were made over all 99 bases for each of the 32 different regions. The calls were compared with the sequences expected for each
35 of the 32 known wild type sequences (see Figs. 5 and 6). For each, the results of the base calling were listed, and the output was sorted based on the number of calls (# correct) that match the expected sequence in that region. In all the

cases to date, when known sequences have been hybridized, the correct sequence was at the top of the list. Additionally, the software gives a detailed assessment of the base calls in each of the four different regions for the top five sequences in the list, giving a clearer picture of where the similarities and differences occur. Result of hybridization between the array and DNA from hBMP-4 and hBMP-5 is shown in Fig. 6.

Diagnosis of a disease is made by obtaining a sample of bodily fluids, tissue, or other nucleic-acid containing material and determines whether a sequence signature present in a possible pathogen or set of pathogens is present. For example, one determines whether a sequence signature of a mycobacterium is present using the methods taught on PCT publication no. W097/29212. If no mycobacterium is present, the presence of another suspected pathogen is explored.

The present invention provides a novel method for performing assays on polynucleotide arrays. While specific examples have been provided, the above description is illustrative and not restrictive. Many variations of the invention will become apparent to those of skill in the art upon review of this specification. The scope of the invention should, therefore, be determined not with reference to the above description, but instead should be determined with reference to the appended claims along with their full scope of equivalents.

All publications and patent documents cited in this application are incorporated by reference in their entirety for all purposes to the same extent as if each individual publication or patent document were so individually denoted.

WHAT IS CLAIMED IS:

1 1. A method for determining whether a target
2 molecule has a sequence from a gene family member comprising:
3 providing an polynucleotide array comprising,
4 for each of at least two different gene family members, a set
5 of polynucleotide probes that define a reference nucleotide
6 sequence from the gene family member;
7 generating hybridization data by performing a
8 hybridization reaction between the target nucleic acid
9 molecule and the probes in the sets and detecting
10 hybridization between the target nucleic acid molecule and
11 each of the probes in the sets; and
12 processing the hybridization data to determine
13 whether the target nucleic acid has the reference sequence
14 from one of the gene family members.

1 2. The method of claim 1 further comprising the
2 step of selecting the target nucleic acid molecule by
3 determining whether the target hybridizes to a nucleic acid
4 probe that hybridizes to a gene encoding the gene family
5 members.

1 3. The method of claim 1 wherein the step of
2 processing is performed by a programmable digital computer.

1 4. The method of claim 1 wherein the
2 polynucleotide array further comprises, for each of the gene
3 family members, a probe set defining a highly conserved region
4 of the gene and a probe set defining a highly variable region
5 of the gene.

1 5. The method of claim 1 wherein the
2 polynucleotide array further comprises, for each of the gene
3 family members, probe sets defining at least two highly
4 conserved regions of the gene and probe sets defining at least
5 two highly variable regions of the gene.

1 6. The method of claim 1 wherein the region codes
2 for amino acid sequence and the array further comprises probe
3 sets defining the degenerate set of nucleotide sequences
4 encoding the amino acid sequence.

1 7. The method of claim 1 further comprising the
2 step of determining the nucleotide sequence of the target
3 nucleic acid molecule if the target does not have the sequence
4 of the region of a gene family member.

1 8. A method of determining whether a nucleic acid
2 in a sample is a member of a gene family, comprising:
3 selecting a hierarchy of assay techniques
4 comprising at least a first and second assay, said first assay
5 being selected to provide a determination of a presence,
6 absence, or variant of a first sequence signature and said
7 second assay being selected to provide a determination of a
8 presence, absence or variant of a second sequence signature;
9 wherein at least one of said assays employs a high-density
10 nucleic acid array;
11 analyzing said nucleic acid sample using said
12 first assay; and
13 determining whether said nucleic acid is a
14 member of said gene family based on the results of said first
15 and second assays.

1 9. The method of claim 8, wherein said first
2 sequence signature is a highly conserved region of a gene
3 family.

1 10. The method of claim 8, wherein said second
2 sequence signature is a non-conserved region of a gene family.

1 11. The method of claim 8, further comprising
2 determining the full length sequence of said nucleic acid
3 sample.

1 12. The method of claim 8, wherein said gene family
2 is the TGF-beta family.

1 13. The method of claim 8, wherein said first or
2 second sequence signature is between 10 and 100 nucleotides in
3 length.

1 14. The method of claim 13, wherein said first or
2 second sequence signature is between 18 and 30 nucleotides in
3 length.

1 15. A method of selecting clones for analysis
2 comprising:
3 providing a support having a variety of clones
4 associated therewith;
5 exposing said support to one or more
6 polynucleotides under low, medium, or high stringency
7 conditions to permit at least some hybridization between said
8 clones and said polynucleotides;
9 identifying said clones that hybridize with
10 said polynucleotides ; and
11 selecting at least one of said clones not
12 identified in said identifying step for analysis.

1 16. The method of claim 15, wherein said support is
2 a high-density nucleic acid array.

1 17. A method of narrowing a sample for analysis,
2 comprising,
3 providing a sample containing nucleic acids;
4 analyzing whether said sample contains a
5 sequence signature using a high-density nucleic acid array;
6 and
7 further analyzing said nucleic acid sample only
8 if said sequence signature is not present.

1 18. A high density nucleic acid array comprising
2 sequence signatures from the TGF-beta gene family.

FIG. 1A

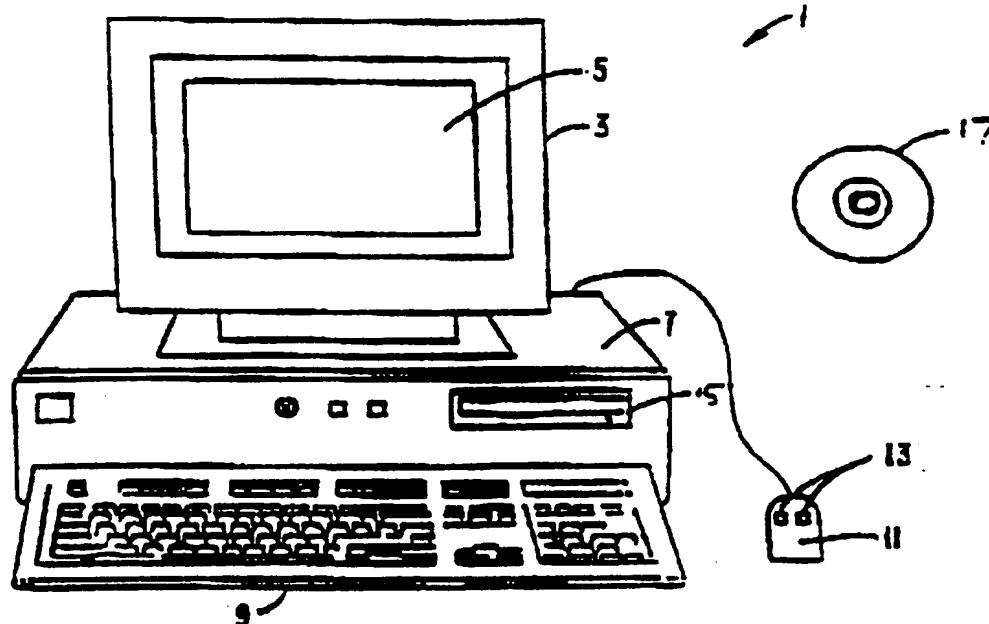
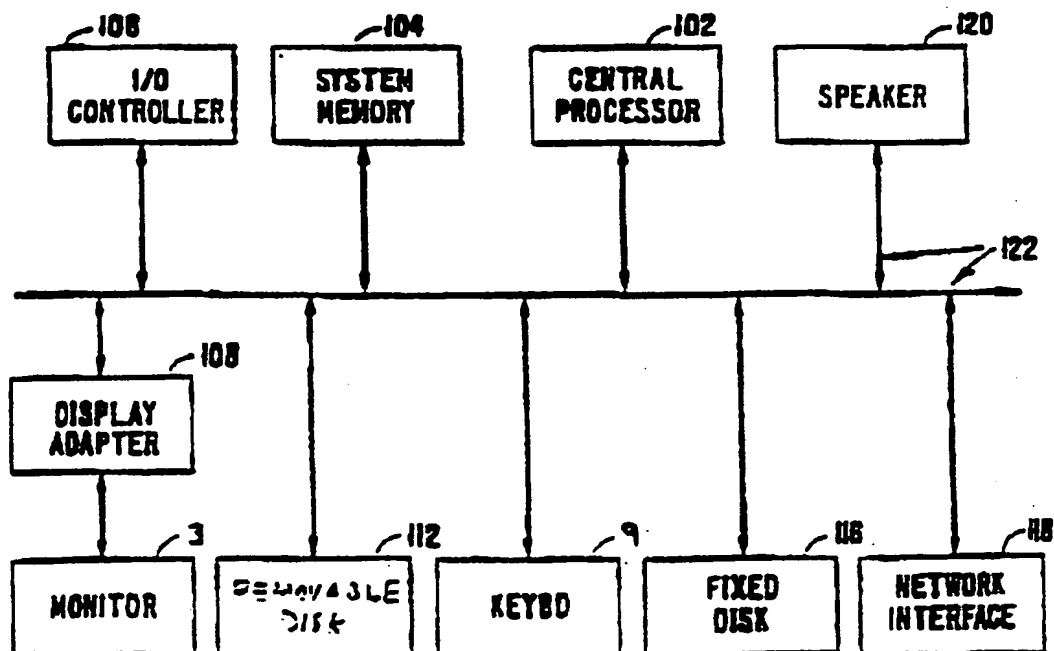


FIG. 1B



File 2

ASN GLY LEU ALA MET ..

SEQUENCE SIGNATURE

ATT GGC AAG GCT ATG

REFERENCE SEQUENCE

a T T G G C
 b T T G G C A
 c T G G C A A
 d G G C A A A
 e G C A A A G
 ...
 e_n

PROBE SET DEFINING
REFERENCE SEQUENCE
(STANDARD TILING)

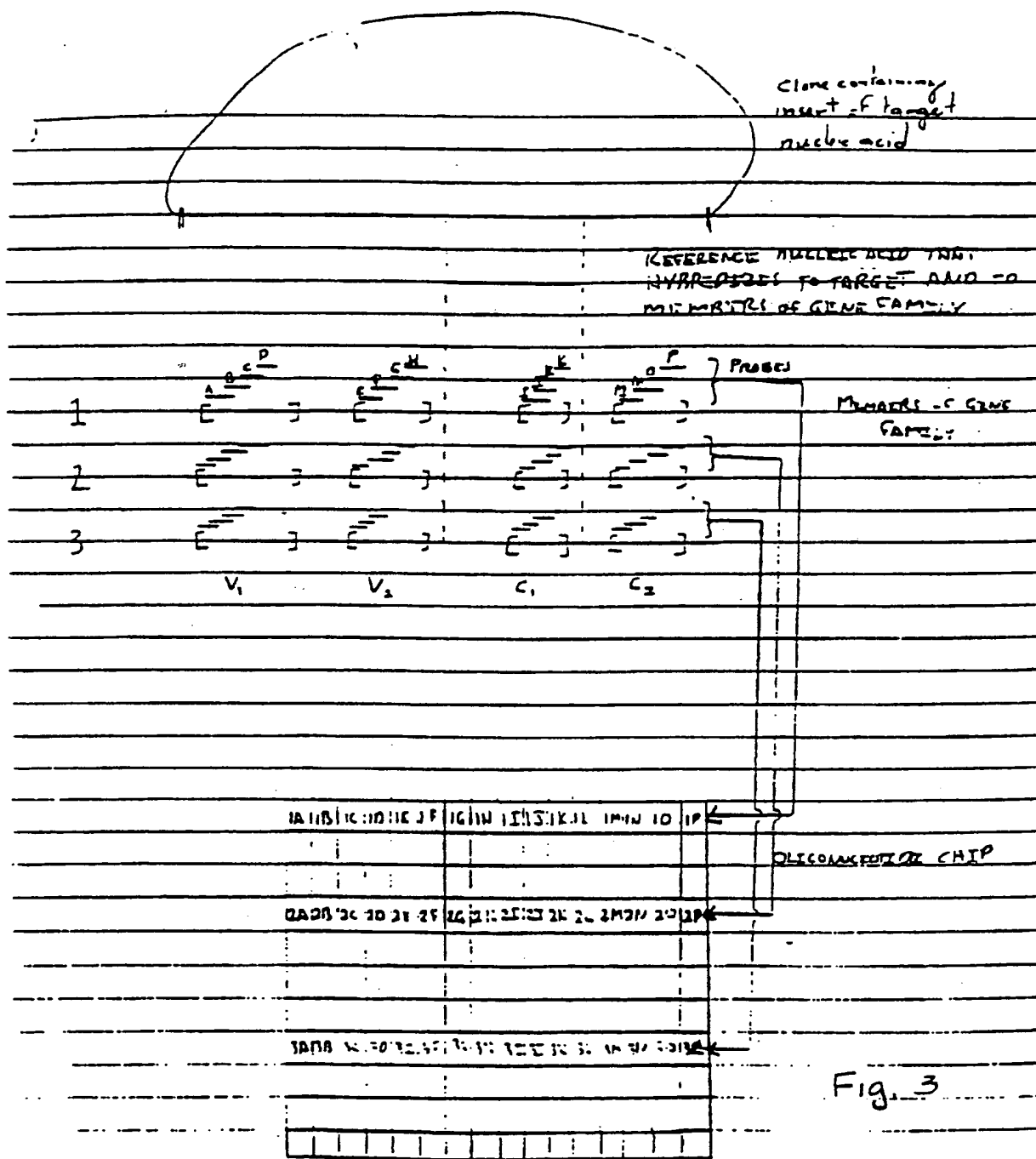
AAC GGA ANG SCA ATG

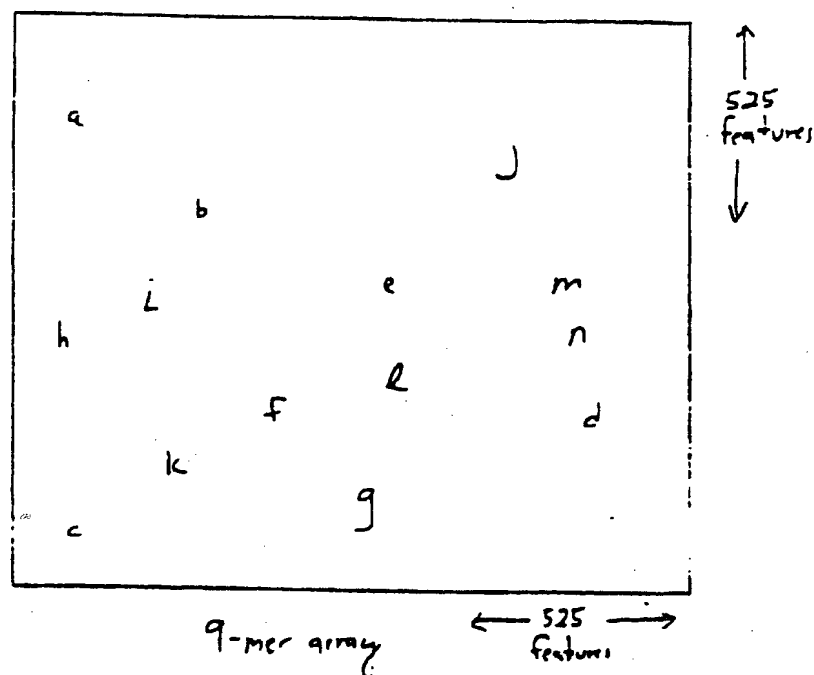
DEGENERATE REFERENCE
SEQUENCE

[AAC	GGA		
	AC	GGA	A	
	C	GGA	AA	
		GGA	AAG	
		GA	AAG	G

PROBE SET DEFINING
REFERENCE SEQUENCE
(STANDARD TILING)

[illegible]





POLYPEPTIDE SIGNATURE SEQUENCE 1

ASN GLY LYS ALA MET

REFERENCE SEQUENCE

ATT GGC AAA GCT ATG

POLYPEPTIDE SIGNATURE SEQUENCE 2

ARG ARG GLY SER PHE

REFERENCE SEQUENCE

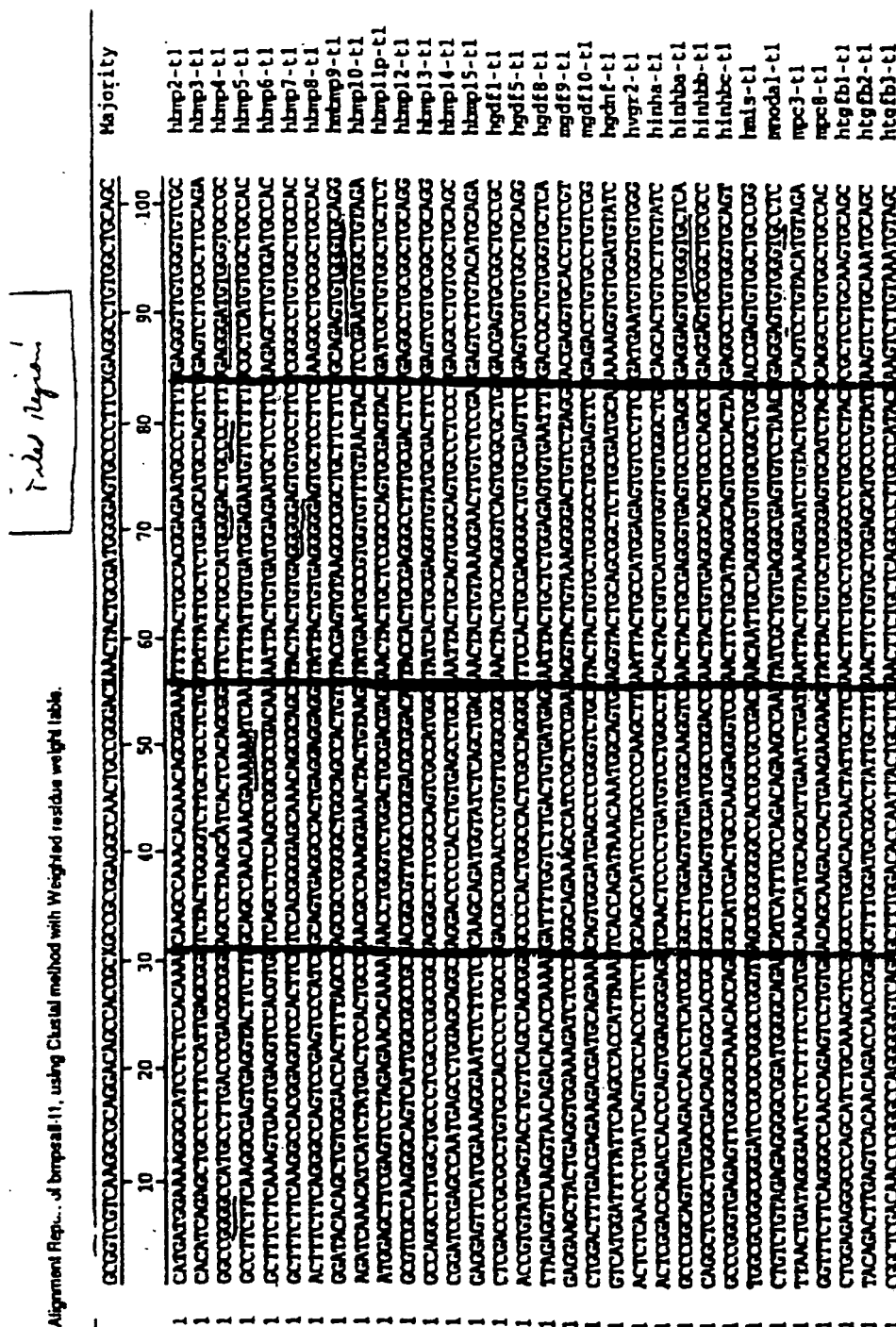
CGC CGC GGA AGT TTT

a	ATT	GGC	AAA	
b	TT	GGC	AAA	G
c	T	GGC	AAA	GC
d		GGC	AAA	GCT
e		GC	AAA	GCT A
f		G	AAA	GCT AT
g		AAA	GCT	ATG

h	CGC	CGC	GGA	
i	GC	CGC	GGA	A
j	C	CGC	GGA	AG
k	CGC	GGA	AGT	
l	GC	GGA	AGT	T
m	CGG	A	AGT	TT
n	GGC	AGT	TTT	

PROBE SETS
 DETIMING
 REFERENCE
 SEQUENCES

Fig. 4



30	24	27	18
just 5' of mature region	Final 8 AA's of mature region	CXCKC Apelin conserved motif	CXCKC region conserved motif end of mature region
5' of <u>1258A</u> cleavage site			
Variable	5		

Fig 5

Sama BMP-7 RNA RESULTS

TITLE	NAME	NG000413	SEMP, BMP-7	0 W	0 DISE	0 Dose	Mean	Median
		0	Correct	0	0	0	866.3	686.6
3m Pi	SEM1	24	24.2	81	13	7	889.3	383.6
	SEM2	24	24.2	39	28	14	325.8	178.6
	SEM3	20	19.2	37	22	10	346.6	230.1
	SEM4	19	19.2	39	22	6	477.4	292.3
	SEM5	18	18.2	39	20	8	323.6	206.9
	SEM6	17	17.2	80	21	2	352.4	213.1
	SEM7	16	16.2	57	21	2	484.3	235.9
	SEM8	15	15.2	41	20	9	268.3	216.9
	SEM9	15	15.2	84	16	11	189.5	168.1
	SEM10	13	12.1	86	21	5	431.2	283.1
	SEM11	12	12.1	49	20	16	248.3	184.7
	SEM12	12	12.1	39	26	11	311.7	189.8
	SEM13	11	11.1	63	14	16	383.2	178.9
	SEM14	11	11.1	48	24	9	218.8	184.1
	SEM15	11	11.1	64	17	5	185.9	189.7
	SEM16	10	10.1	70	14	4	288.3	186.4
	SEM17	10	10.1	72	11	5	320.8	285.9
	SEM18	9	9.1	76	18	4	258.3	184.8
	SEM19	9	9.1	88	25	7	330.8	178.1
	SEM20	9	9.1	88	27	4	194.4	170.6
	SEM21	7	7.1	88	22	11	218.8	176.1
	SEM22	6	6.1	87	18	8	188.8	167.9
	SEM23	6	6.1	63	26	4	348.5	212.1
	SEM24	5	5.1	68	24	3	178.9	183.2
	SEM25	5	5.1	78	17	2	188.9	185.9
	SEM26	4	4.0	86	6	3	217.6	185.8
	SEM27	3	3.0	77	21	6	168.7	188.1
	SEM28	3	3.0	71	16	7	214.8	188.6
	SEM29	2	2.0	80	11	3	282.7	288.6
	SEM30	2	2.0	70	18	8	198.1	189.7
	SEM31	2	2.0	72	16	5		
	SEM32	0	0.0	88	13	1		

SD = 60

Anti-sense BMP-5 RNA

NAME	NG000415	SEMP, BMP-5	0 W	0 DISE	0 Dose	Mean	Median
	0	Correct	0	0	0	1838.4	1838.2
SEM1	27	26.8	33	27	4	318.9	182.2
SEM2	23	22.2	33	21	11	307.1	169.1
SEM3	20	18.2	46	21	7	188.9	119.9
SEM4	17	17.2	48	23	7	234.4	118.4
SEM5	16	16.2	31	24	13	182.6	122.6
SEM6	15	15.2	67	20	5	178.4	185.7
SEM7	15	15.2	49	17	7	168.6	148.4
SEM8	14	14.1	38	22	15	215.4	125.7
SEM9	12	12.1	49	22	18	336.8	128.2
SEM10	10	10.1	80	15	6	248.3	138.7
SEM11	9	9.1	61	25	17	127.2	101.8
SEM12	9	9.1	67	13	4	188.6	107.6
SEM13	9	9.1	86	18	9	185.2	189.8
SEM14	8	8.1	33	29	6	166.1	98.2
SEM15	7	7.1	80	23	7	278.1	127.8
SEM16	6	6.1	40	27	11	136.8	109.8
SEM17	6	6.1	41	23	3	168.2	112.8
SEM18	6	6.1	33	28	7	182.9	111.8
SEM19	5	5.1	68	19	7	324.6	189.1
SEM20	5	5.1	3	24	14	183.8	98.2
SEM21	5	5.1	54	23	9	274.8	182.1
SEM22	5	5.1	84	27	5	124.1	95.8
SEM23	5	5.1	64	18	8	165.4	108.6
SEM24	3	3.1	83	21	18	327.2	189.7
SEM25	4	4.0	68	14	3	101.2	91.8
SEM26	4	4.0	43	18	6	172.4	107.6
SEM27	3	3.0	57	15	9	118.7	91.4
SEM28	3	3.0	43	19	7	162.6	189.8
SEM29	3	3.0	69	23	2	182.2	94.1
SEM30	2	2.0	76	12	4	154.6	118.9
SEM31	1	1.0	74	14	5	122.4	108.1

SD = 30

Fig. 6

INTERNATIONAL SEARCH REPORT

International application No.
US97/17002

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) : C12Q 1/68; C07H 21/04; G06F 17/00

US CL : 435/6; 536/24.3; 364/400

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6; 536/24.3; 364/400

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
NONE

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS, MEDLINE, BIOSIS, CAPLUS

search terms: array, probes, nucleic acid, hybridization, computer digital

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	WO 89/10977 A1 (ISIS INNOVATION) 16 November 1989 (16-11-89), see entire document.	1-15

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents	* T* later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
* A* document defining the general state of the art which is not considered to be of particular relevance	* X* document of particular relevance, the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
* B* earlier document published on or after the international filing date	* Y* document of particular relevance, the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
* L* document which may throw doubts on priority claimed or which is cited to establish the publication date of another citation or other special reason (as specified)	* A* document member of the same patent family
* O* document referring to an oral disclosure, use, exhibition or other means	
* P* document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search

01 DECEMBER 1997

Date of mailing of the international search report

31 DEC 1997

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231

Facsimile No. (703) 305-3230

Authorized officer

JOYCE TUNG

Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet)(July 1992)*

